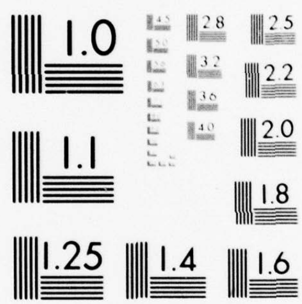
AD-A066 161 SPEECH COMMUNICATIONS RESEARCH LAB LOS ANGELES CA F/G 17/2
REVIEW OF THE ARPA SUR PROJECT AND SURVEY OF CURRENT TECHNOLOGY--ETC(U)
JAN 79 W A LEA, J E SHOUP N00014-77-C-0570

UNCLASSIFIED NL

1 OF 2

AD
A066161



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

AD A0 66161

DDC FILE COPY

REVIEW OF THE ARPA SUR PROJECT AND

REVIEW OF THE ARPA SUR PROJECT AND SURVEY OF CURRENT TECHNOLOGY IN SPEECH UNDERSTANDING

Wayne A. Lea

June E. Shoup

Speech Communications Research Laboratory
806 West Adams Boulevard
Los Angeles, CA. 90007

FINAL REPORT
OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-77-C-0570

Submitted to

Office of Naval Research
Department of Navy
800 North Quincy Street
Arlington, Virginia 22217

JANUARY 16, 1979

This work was supported by the Office of Naval Research, the Air Force Office of Scientific Research, and the Advanced Research Projects Agency and monitored by the Office of Naval Research under contract number N00014-77-C-0570. The views and conclusions expressed in this document are those of the authors alone (or of the experts they polled, whenever so indicated), and should not be interpreted as necessarily representing the official policies of the sponsors or monitoring agency, or the U.S. Government.

(10)

REVIEW OF THE ARPA SUR PROJECT AND SURVEY OF CURRENT TECHNOLOGY IN SPEECH UNDERSTANDING

Wayne A. Lea
June E. Shoup

Speech Communications Research Laboratory
806 West Adams Boulevard
Los Angeles, CA. 90007

FINAL REPORT
OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-77-C-0570

submitted to

Office of Naval Research
Department of Navy
800 North Quincy Street
Arlington, Virgina 22217

January 16, 1979

This work was supported by the Office of Naval Research, the Air Force Office of Scientific Research, and the Advanced Research Projects Agency and monitored by the Office of Naval Research under contract number N00014-77-C-0570. The views and conclusions expressed in this document are those of the authors alone (or of the experts they have polled, whenever so indicated), and should not be interpreted as necessarily representing the official policies of the sponsors or monitoring agency, or the U.S. Government.

TABLE OF CONTENTS

PREFACE

This report summarizes the major technical accomplishments of a recent project which was conducted to demonstrate the technical feasibility of naturally speaking commands to a computer, and having the machine "understand" the content of the spoken sentence sufficiently to produce a correct response. The purposes, history, and current capability in voice input to machines are briefly presented, to provide the context in which accomplishments can be evaluated, and to provide the background for the authors' conclusions about future work that still needs to be done. There are many technical terms and detailed concepts in the report which might be foreign to readers whose expertise does not lie in speech sciences, computer technology, or engineering. Consequently, we offer in this preface an introductory overview and a non-technical explanation of the origin and current directions of this work.

1. What is a Speech Understanding System?

A speech understanding system of the most general kind would be a computer system that could (1) accept verbal commands and questions spoken into a microphone by any trained user, (2) figure out the intended meaning of the sentence, and (3) generate an appropriate response. The technical difficulties associated with the design of a system this powerful are well beyond the present and near-future capabilities of computer scientists, speech scientists, linguists, and engineers.

However, recent research to be described in this report suggests that limited speech understanding systems now constitute a practical engineering goal. A limited system attempts to recognize sentences constructed from small to moderate vocabularies, but with severe constraints on acceptable word order. A number of technical problems remain to be solved before practical systems become readily available, but a program of research is described in the body of this report that can lead to the development of appropriate solutions.

2. Why Build Limited Speech Understanding Systems?

Is limited speech understanding a desirable goal in terms of potential applications and in terms of the long-range implications for society as a whole? It is concluded that important applications do exist and, more importantly, that spoken input capabilities are likely to make computers more accessible and useful to a wider segment of the general public. A critical problem in the effective use of machines concerns the ease and accuracy with which commands are communicated. This need for the human user to effectively instruct the machine is clearly evident in large control consoles that are used with navigation systems for ships and aircraft and with other complex command and control systems. However, effective communication of commands is also needed for the best use of small data-entry devices with only a few knobs or switches. Errors in pushing the wrong button, dialing or tuning incorrectly, or using the wrong sequence of commands can be wasteful of time and effort, frustrating to personnel, and sometimes directly hazardous. In addition to being prone to errors, the unnatural ways of communicating can be tedious, and expensive (in that they require highly trained personnel). Conversing with a machine in spoken form is clearly one of the most natural means of instructing a machine. For years, there has been a growing interest in versatile, "natural-to-

the-human" techniques for human communication to computers and other machines. Computer specialists have been developing a variety of devices, programs, and "natural" programming languages for rapid interaction between humans and machines.

Speech input to machines offers an unprecedented mobility to computer users, in that the user need not be in actual physical contact with the machine, but rather may walk around, turn aside, and, most importantly, use his hands and eyes for other tasks while speaking instructions to the machine. Speech is fast, spontaneous, and nearly universal among humans. Indeed, experiments with humans interacting to cooperatively solve various problems have shown that people can find solutions twice as fast when they are allowed to converse in spoken form, in contrast to when they communicate by typewriter, handwriting, or visual signs.

Examples of applications where spoken input capabilities are clearly important include (1) situations where the user's arms are occupied by other tasks (e.g. air traffic control, sorting, and assembly), or where the user must move about, (2) situations requiring remote access (by ordinary telephone) to computerized information systems, banking systems, and systems for performing commercial transactions, (3) instructional situations of many kinds, but particularly those concerned with language development and training in skilled communications (e.g. air traffic control), and (4) situations where the computer acts as an expert consultant on some construction/repair topic.

4. Why is Speech Understanding Hard for a Computer?

Part of what makes the development of speech recognizers so hard is that different people speak in different ways. Also, a noisy acoustic environment may interfere with reliable interpretation of the acoustic speech signal. In addition, even the same single talker will vary from time to time in his pronunciations. The problem is complicated considerably by the complexities of naturally flowing connected speech.

A good way to examine the difficulties involved in speech understanding by computer is to compare speech understanding with isolated word recognition. There has been considerable success in designing and marketing computer systems that recognize an isolated word from a set of 10 to as many as 1000 alternatives. The best performance is achieved by systems that treat each word as a spectral pattern, and require each user to speak all words into the computer so as to generate a personal set of word patterns.

It might seem that this technology could be directly applied to the recognition of words in spoken sentences, but this is unfortunately not the case. There are no pauses between words of fluent speech. The speech production organs (tongue, lips, larynx, etc.) move continuously from one articulation to the next. This changes the acoustic characteristics of the sounds at the beginnings and ends of adjoining words, and can completely obscure the locations of word boundaries. There are also special rules that speakers of a given dialect use to simplify word pronunciations in some contexts. For example, "you" is pronounced normally in "are you going?", but "you" is pronounced as "ja" in "would ja please pass the butter". In this case, the "d" from the previous word causes the change. While one could imagine solving this problem by a brute force storage of each person's way of speaking each total sentence, it is impractical to collect and store examples of every sentence that is to be understood, or to compare any large set of such stored sentences with an incoming utterance.

It is clear that human listeners use expectations all of the time to help figure out what is being said. For a computer to simulate this ability, it would

be necessary to include considerable knowledge about the world and the nature of what people tend to say in various contexts. Thus it is not surprising that general speech understanding is an impractical goal, and many artificial constraints must be imposed to help the computer deduce what has been said.

A speech understanding system can be thought of as a set of interacting knowledge sources, each contributing to the interpretation of an unknown input waveform. Components of a limited speech understanding system might perform functions such as:

- extracting acoustic parameters from the speech waveform ("acoustic analysis")

- identifying the vowels and consonants that are present ("phonetic analysis")

- comparing this sequence of speech sounds with expected pronunciations of words ("word matching")

- testing whether a hypothesized word is syntactically consistent with words already recognized, and using syntactic constraints to predict likely upcoming words ("syntactic analysis")

- testing the meaningfulness of hypothesized word sequences ("semantic analysis")

- predicting likely future words based on prior discourse and on the task being performed ("pragmatic analysis")

The problem is that none of these components currently can be made to perform reliably. A strategy must be found to combine information from all components in such a way that errors in intermediate stages of analysis do not result in errors of sentence understanding.

4. Background and Objectives of the ARPA SUR Project

The idea of talking to machines is not exactly new. For about three decades, work was done under the major simplifications provided by confining spoken inputs to single words spoken in isolation, with easily detected pauses before and after each word. Most work on isolated word recognizers has been confined to speaker-dependent systems, for which each new talker must "train" the system to recognize his particular way of saying the alternative words.

Early work on speech recognition, conducted in the 1950's, made use of then current technology for decomposing any mathematical signal (such as an acoustic wave, or light, or radio waves) into simpler component waveforms (called "sine waves") that smoothly varied in a periodic manner, from large values to small values, and back to large values, etc. These recurrent waves were said to have a "frequency" which reflected the number of full cyclic vibrations occurring per unit time. Speech could be decomposed into waves of different frequencies, by mathematical processes called "Fourier analysis" or "frequency spectrum analysis". Such ideal mathematical decompositions of signals could be approximated by electronic "filters" which separated different frequencies from each other. Thus, frequency spectrum analysis could be accomplished by available electronic hardware.

Speech had long been acknowledged to be produced by the human vocal mechanism in such a way as to have many frequencies in it, with some frequencies highly accented by the mouth and throat acting like a specially-shaped acoustic tube. Such an acoustic tube has "resonances" that are related to the geometry of the tube; these natural resonances of the human vocal system are commonly called "formants". Other important aspects of the speech that were expected to be important in speech recognition included the amplitude or loudness of the speech and the "pitch" of the voice (well represented by different musical notes on which a particular vowel like "ah" could be sung or said). Some speech sounds (for example, consonants "s", "k", and "p" in a word like "skip") are spoken without vibration of the talker's vocal cords, and are called "unvoiced" sounds, while others (like vowels, or some consonants like "m", "n", and "r") involve vocal cord vibration and are called "voiced". Other classes of speech sounds (called "phoneme classes") were distinguishable, like the hissing "fricatives" ("s", "f", etc.), the silent-or-near-silent-then-exploding "stop consonants" ("p", "t", "k", "b", "d", "g"), and "nasals" (which involve airflow out through the nose like in "n" or "m").

From the early stages of speech recognition, controversies arose about the utility of all these speech categories for the practical processes of machine recognition of speech. Some argued that we need to understand how the human produces speech, and use distinctions or categories ("phonemes") that the human uses in producing or perceiving various vowels and consonants. Other more engineering-oriented or mathematically-inclined workers suggested that direct comparisons or "correlations" should be done between input signals and stored "templates" or exemplars of the expected words, as obtained from previous "training samples" (sample utterances of the words as spoken earlier and stored away for comparison).

Early word recognizers worked fairly well with simple correlations of the total pattern for a word with stored templates, and later work showed only slight improvements due to techniques that detected the vowels and consonants that comprised the word. Around twenty years ago, studies showed that recognition accuracy could be improved if the speech to be recognized was "time normalized", so it was stretched or shrunk to the duration of the stored templates. About that same time, the digital computer was first used in speech recognition. (Before that time, work had been done with special-purpose electronic hardware).

Throughout the 1960's, continued work was done on both the mathematical and the phonemic approaches to recognition, and initial attempts were made to recognize "continuous speech", such as word sequences without pauses between words. Important advancements were made in the classification of vowels and consonants occurring in continuous speech. Also, major strides were made in detailed mathematical procedures for signal analysis, such as (a) the "fast Fourier transform" (which permitted rapid determination of the frequency spectrum in each short region of the speech, using a digital computer), and (b) "Linear predictive analysis" (which separated the effects of the talker's vocal cords and his mouth-throat acoustic tube, so that formants and other interesting features of the speech signal could be more readily and accurately detected).

The 1960's also saw the growth of an almost-universal call for the use of "higher-level linguistic analysis" in speech recognizers, so that expected and grammatically-acceptable sequences of words would be used to limit the possible words that might be guessed to occur at various points throughout the utterance. Also, "semantic" constraints on meaningful sequences of words could be used to

rule out some word sequences that might be incorrectly hypothesized for the speech. One might also use known regularities of English sound structure or "phonology" to select the most likely words being pronounced. The basic contention was that as one moves from simple isolated words to the handling of continuously spoken sentences, all kinds of confusions in possible wording might arise, and the linguistic knowledge would help keep the alternative word sequences to be tried down to a minimum. This incorporation of linguistic knowledge seemed to be a formidable task, and at the end of that decade some influential researchers were pessimistic about the foreseeable future producing any adequate recognizers of continuous speech.

Other major advances were being made however, in computer technology and "artificial intelligence" (the ability to perform tasks on machines that would be said to involve "intelligence" if humans did them). Procedures had been developed to have computers play games like checkers and chess, to make logical deductions and inferences, to recognize patterns such as handwritten characters or tanks in camouflage, and to rapidly search among thousands of alternatives to find the best solution to a problem. In addition, what have sometimes been called "friendly systems" were developed, which permitted users to very readily and naturally interact with a computer, without the need for cumbersome mathematical languages or unnatural diversions into the intricate details of inner workings of the machine. "Time sharing systems" permitted more than one user to effectively use a machine at the same time, and opened up the possibility for large groups of researchers to cooperatively work on various aspects of a complex problem (such as spoken sentence recognition), using the same computer system. The stage had clearly been set for major advances on the complex problem of continuous speech recognition.

The Advanced Research Projects Agency (ARPA) apparently recognized this coalescing of all the essential ingredients for an effective assault on the task of understanding spoken sentences. ARPA had been influential in funding much of the relevant advances in artificial intelligence and advanced computer technology, and could see the prospects for applying that work, and other recent advances in speech sciences and linguistic processing, to the recognition of speech. There apparently also was a keen awareness of the value of using speech understanding as a task which involved many sources of incomplete knowledge, each working together to help refine decisions about the content of an utterance. Systems could exploit recent advances in "syntactic parsing" (the determination of the grammatical structure or phrasal groupings in a sentence), "semantic analysis" (the interpretation of the "meaning" of a sentence), and "pragmatic analysis" (the determination of the appropriateness of a particular sentence in the context of previous discourse and in accord with the constraints of the task being performed by the human-plus-machine interactive system). The emerging theories of the sound structure of English ("phonological rules") and the initial attempts at characterizing the intonation, timing, rhythm, and accentual patterns of speech (so-called "prosodic structures") could be incorporated into the system. And all this could be coupled with the advancing techniques in acoustic analysis and vowel and consonant identification. Alternative "control structures" existed for integrating all these processes into one cohesive system that carefully focussed on the most promising information and properly scheduled and coordinated all the subprocesses.

In 1971, the Advanced Research Projects Agency (ARPA) of the United States Department of Defense, commissioned a study group to explore the design of a large project for determining the feasibility of systems that understand speech. The conditions seemed ripe then for integrating many disciplines into one cohesive effort. A dominent force in that study group was a collection of artificial

intelligence experts who had been effective in previous ARPA projects dealing with other artificial intelligence tasks. This group defined an ambitious five year "Speech Understanding Research" (SUR) project, involving five initial contractors who were to build speech understanding systems. At the mid-term of the project, each contractor's intermediate test system was to be evaluated, and the best three or four systems were to be continued in development. A comprehensive set of system specifications were defined for the final evaluation of the resulting systems. These specifications did not necessarily represent "the last word" in necessary system performance conditions, but they were the best estimate of the study group concerning reasonable goals that would show the feasibility of speech understanding and would signal the emergence of a promising overall technology for the comprehension of continuously spoken sentences by complex systems. The primary initial goal (and the longest-lasting legacy) of this project was the successful demonstration of an emerging interdisciplinary technology for effective machine comprehension of continuously spoken sentences. Many problems in speech understanding were uncovered, and some promising initial solutions developed, but much more work is still to be done. To understand this, we need to look more closely at the goals and accomplishments of the ARPA SUR project, the overall current status of speech recognition technology, and the problem areas or "gaps" in technology that remain.

5. Goals and Accomplishments of the ARPA SUR Project

The ambitious system specifications of the ARPA SUR project called for machines that would accurately (i.e., for over 90% of the correctly-spoken sentences) accept continuous speech from many cooperative speakers, with near-ideal conditions of quiet rooms and high-fidelity equipment. Sentences were to be highly-stylized structures defined by a small grammar, using a 1000-word vocabulary. Realizing both the complexity of the problem and the prospects for rapid advances in computer technology, they called for the recognition to be accomplished on very large fast computers that could handle about 100 million internal instructions per second (which is about 100 or more times as powerful as the actual computers the systems were finally built on), and yet they allowed the computer processing to take several times as much time as the duration of the spoken sentence. In computer parlance, the processing then requires "several times real time".

In addition to five original system-building contractors (who tried alternative system designs), the project included four research contractors who were charged with developing advanced ideas for improving the recognition techniques. At the mid term of the project, the best intermediate systems were selected for continued development. In the fall of 1976, this largest project ever undertaken in machine understanding of speech came to an end with the demonstration of several systems that could understand spoken sentences. Carnegie-Mellon University demonstrated two alternative system designs (called "Harpy" and "Hearsay II"), Bolt Beranek and Newman, Inc. demonstrated the "Hear What I Mean" (HWIM) system, and System Development Corporation demonstrated a system. One of the systems, the Harpy system developed at Carnegie-Mellon University, basically met or exceeded the system goals by correctly understanding 95% of the sentences spoken by five talkers, using a 1011-word vocabulary and a highly-constrained grammar of sentences relevant to a task concerning the retrieval of documents from the computer memory. Five talkers is not "many", and the tests were done on only a small set of 184 sentences, due to time and money limitations. However, the system did work well even when the original specifications were <u>exceeded</u> by having it handle somewhat noisy speech with inexpensive (lower-fidelity) michrophones. Harpy not only met the

"letter of the law" by matching the ambitious goals for the project; it also fulfilled the "spirit" of the project by demonstrating the feasibility of a limited (but potentially useful) technology for computer understanding of continuously-spoken sentences. Also, in line with the spirit of the project, it made effective use of strict constraints on allowable (grammatical, meaningful, and relevant) word sequences, to bring the task within manageable limits. Other final ARPA SUR demonstration systems had higher error rates primarily because they dealt with more difficult tasks, used more general techniques that could have been used for additional more ambitious tasks, and were not as carefully tested and adjusted as Harpy before the final demonstration tests. The Harpy system also benefited from the discovery of a new way to incorporate knowledge about the acoustic properties of speech sounds and the phonetic composition of words into a special network. The network structure permitted very rapid examination of many alternative word sequences before selecting the best-scoring word sequence as the sentence spoken by the user.

6. Evaluating the ARPA SUR Contributions

The successful attainment of the original ambitious system goals was a major contribution of the ARPA SUR project, and the Harpy system performance now provides a baseline or benchmark for assessing future work. However, many other important contributions were made, as is detailed in this report. Most of these valuable contributions are highly technical, but we can make several general observations about ARPA SUR contributions. Of major importance is that we now know that continuous speech can be accurately recognized, at least for the case of sentences related to a limited task. What's more, the original premise that recognition accuracy would be aided by judicious use of linguistic constraints has been vividly demonstrated. The value of artificial intelligence ideas like efficient search strategies and cooperation among several incomplete sources of knowledge has been shown. Several promising alternative system structures have been tested, and major advances were made in certain system components such as vowel and consonant recognition schemes, phonological rules, prosodic analysis routines, and word identification procedures.

Thus, the ARPA SUR project produced major strides in the necessary technology for commanding machines by naturally spoken sentences.

7. Scientific Problems that Remain

Despite these important and satisfying advances, the ARPA SUR project and other recent work have demonstrated that in almost every component or aspect of a recognition system, there still is need for further major improvements. The problem of voice input is not solved. No system currently can precisely and correctly identify much more than half of the vowels and consonants in continuous speech, yet experiments conducted during and following the ARPA SUR project suggest that humans can do much better than these current vowel and consonant identification schemes. Recognizers have not even incorporated or adequately tested many of the published rules concerning English phonological structure (the allowable sequences of vowels and consonants, the effects of one sound on its neighbors and vice versa, and the effects at boundaries between words). Prosodic structures (intonation, stress patterns, and timing of speech events) show great promise of aiding word recognition and detection of several aspects of grammatical structure, but prosodic information has had virtually no impact on the performance of previous recognizers. While several promising techniques have previously been developed for identifying words by

their resemblences to expected pronunciations, further work is still needed to increase the accuracy of word matching. At the higher levels of linguistic analysis (dealing with larger units like phrases and sentences), efficient constraints must be developed that still allow future expansions to more difficult tasks. We need precise methods for assessing the relative complexities of various recognition tasks and for adequately evaluating the total performance of a recognizer.

The systems developed under the ARPA project were specifically intended to demonstrate feasibility, not practicality. Thus the choice of grammar and task domain was dictated by experimental convenience and not by considerations of potential applications. Similarly, algorithms were developed without consideration of speed of execution or the size and power of the computer that is required.

Now it is time to examine the state of the art, and see what is needed to reach the goal of good performance and low cost in real applications. There are two issues: (1) is the technology there to perform a given task, and (2) can it be performed fast enough and at reasonable cost? The report concludes that the technology is nearly there for very modest applications, but that significant advances are within reach if one engages in specific additional research, taking advantage of what has been learned through the ARPA project and other efforts.

While current technology offers several commercial devices for isolated word recognition, and one system that is purported to accurately handle restricted continuous speech (digit strings or highly constrained word sequences), still there are major "gaps" in current technology that must be filled before speech recognizers will fulfill their potential as versatile tools for conversing with machine. For example, human observers can do quite well, about 90 percent correct, in identifying vowels and consonants of nonsense words placed in spoken sentences. There is no theoretical reason why computers cannot approach this performance, but careful research is needed to reach such a goal. This is probably the research area with the highest potential payoff in terms of improving the performance of limited speech understanding systems. However, improvements are possible in all of the components of speech understanding systems.

It is also clear that strong syntactic constraints are essential given the errorful performance of individual system components, but one has to be careful to devise grammatical constraints that are easy for the user to obey. For example, Harpy was designed to answer questions about newspaper articles, but the grammatical constraints that were imposed resulted in a highly unnatural set of user restrictions.

8. Recommendations

This report offers specific recommendations for advancing the technology of speech understanding. Highlights include proposals to improve and extend Harpy for applications requiring the simplest grammars (for example by expanding the detailed acoustic-phonetic knowledge contained in the Harpy network). Basic research involving more ambitious systems is also needed for many applications. This requires the creation of facilities and research teams that can work on improving specific knowledge sources within the context of a working speech understanding system.

Funding of these and other recommendations contained in the report will not occur spontaneously. The cost and interdisciplinary nature of the research has resulted in a period following the ARPA SUR project where adequate funds are not being spent on speech understanding. We argue that funding is needed, results

are attainable, and a proposed coordinated program of DOD sponsored work deserves immediate attention.

9. Additional Reading

Lea, W.A. (1979) Trends in Speech Recognition. Englewood Cliffs, N.J.: Prentice-Hall.

Klatt, D.H. (1977), "Review of the ARPA Speech Understanding Project", The Journal of the Acoustical Society of America, Vol. 62, 1345-1366.

EXECUTIVE SUMMARY

In the fall of 1976, the largest project ever undertaken in machine recognition of speech was successfully completed, with the feasibility demonstration of several systems that could understand spoken sentences. That 5-year, $15-million project was sponsored by the Advanced Research Projects Agency (ARPA) of the United States Department of Defense (DOD), for "Speech Understanding Research" (hence the name "ARPA SUR project"). A detailed technical review of that project is presented here, along with a survey of the current state of speech recognition technology. However, the review and survey are presented with tomorrow in mind, not yesterdays or fleeting todays, so that past successes and difficulties can lead to recommendations for further work that needs to be done.

The following logically successive questions are addressed, leading towards specific recommendations for further work to be funded in speech recognition:

1. Why is voice input to computers wanted?
2. How much can and should be said to a machine for each application?
3. What work was done prior to the ARPA SUR project?
4. What advances did the ARPA SUR project contribute?
5. What then is the current state of speech recognition technology?
6. What are the remaining "gaps" (system inadequacies, unsolved problems, and scientific and practical needs) in the current technology, and which of those are most significant?
7. What work is recommended for filling those gaps and advancing speech recognition?
8. What mechanisms are suggested for undertaking the recommended work?

Given such recommendations and mechanisms for further work, DOD representatives should have general guidelines for relating this advancing technology to their specific operational applications.

The review and recommendations presented in this report are based in part on a survey of expert opinions obtained from site visits with contractors in the ARPA SUR project, other industrial developers of advanced speech recognition systems, commercial suppliers of recognition devices, and some government sponsors of research and development work in this field. Over 100 speech recognition experts were conferred with, and a detailed questionnaire was completed by 34 respondents. Some of the ideas included here were presented at technical conferences and at an ONR-sponsored workshop for government funders of speech recognition.

We shall now consider each of the eight basic questions mentioned previously, and offer brief answers in the following correspondingly numbered paragraphs.

(1) Regarding the first question of WHY there should be interest in voice input to computers, we may note the following advantages. Speech:

- Is the human's most natural (familiar and spontaneous) communication modality;
- Requires little training (except for how to constrain oneself to saying only what the machine can understand);
- Is the human's fastest modality, requiring less than half the time to perform a task as other modalities do;
- Permits multimodal communication;
- Allows a freedom of movement and orientation;
- Requires little or no instrument panel space in aircraft (and is somewhat insensitive to acceleration and vibrations); and
- Frees the speaker's hands and eyes for other tasks.

However, spoken words, phrases, and sentences do pose some problems, in that speech is:

- Inconsistent from time to time;
- Sensitive to talker differences;
- Not private;
- Subject to interpretation errors due to environmental noise and distortions; and
- Currently fairly expensive compared to other computer-input devices.

Experiments with human interactions, and practical experience with recognition devices, demonstrate that the advantages of speech far outweigh the disadvantages, and the disadvantages generally can be overcome. Applications of limited speech recognizers in many commercial and military installations have clearly shown that speech input reduces workloads, time involved in various tasks, manpower needs, and costs, while often improving worker satisfaction.

(2) A second major point about speech recognition is that there is a SPECTRUM OF POSSIBLE RECOGNITION CAPABILITIES, including:

- Isolated word recognizers, which independently handle words that are preceded and followed by pauses;
- Recognizers of sequences of isolated words, which use sequence constraints ("syntax") to limit what alternative words must be distinguished at each stage in the sequence;
- Word spotting systems, which detect occurrences of key information-carrying words in the context of free-flowing continuous speech;
- Digit string recognizers, which handle uninterrupted sequences of spoken digits;
- Word sequence recognizers, which identify uninterrupted (but strictly formatted) sequences of words;
- Restricted speech understanding systems, which handle total sentences relevant to a specific task; and
- Task-independent continuous speech recognizers, which identify wording of sentences without restriction to a specific task.

For each speech input facility, one must ask which of these recognition capabilities is needed. It is then appropriate to ask what has been learned from previous attempts to develop and use each such type of recognition system.

(3) Throughout the past 26 years, a variety of projects have been undertaken to develop the various speech recognition capabilities, but the primary successes were for the limited problem of isolated word recognition. Basic technical issues were addressed, such as:

- Whether to identify the total (unsegmented) pattern of the word or to segment the word into smaller units like vowels and consonants;
- What distinguishing features of the speech to monitor for identifying words or other units;
- How to time normalize, to adjust for fast versus slow pronunciations;
- How to handle larger vocabulary sizes; and
- What to do about the peculiar problems of continuous speech (such as difficulty of detecting word boundaries, and effects of context on the pronunciation of words).

The HISTORY OF WORK IN SPEECH RECOGNITION was speckled with limited successes and repetitive rediscoveries of old ideas, and yet with a growing ability to successfully handle small or moderate sized vocabularies of isolated words, provided the system was trained to handle the talker's voice. Thus, in 1971, when the ARPA SUR project began, a few prototype isolated word recognizers could correctly identify over 95% of the words spoken by one or a few talkers, when in a fairly quiet environment. Nothing like a 1000 word vocabulary had ever been attacked, and existing examples of continuous speech recognizers were far too limited to be extendable. While there had been a growing call for the use of higher-level linguistic constraints (syntax, semantics, etc.) to limit the alternative words that had to be selected among at each point in an utterance, there still were no "multiple knowledge source" systems that had cooperating knowledge sources for:

- Extracting important acoustic parameters ("acoustic analysis");
- Identifying vowels and consonants in the speech ("phonetic analysis");
- Matching sequences of speech sounds to expected pronunciations of words ("word matching");
- Using stress, intonation, and the timing of speech to identify aspects of the structure of the sentence ("prosodic analysis");
- Verifying the grammaticality of hypothesized word sequences and predicting the possible identities of unidentified words by contextual constraints ("syntactic analysis");
- Testing the meaningfulness of apparent word sequences and hypothesizing other meaningful and semantically related words that might extend partial interpretations of the sentence ("semantic analysis");
- Determining the plausibility of hypothesized word sequences, based on the discourse context and the task being performed ("pragmatic analysis").

Only a couple of basic system structures were known that might integrate all these system components together into an efficient and accurate recognition strategy. Little knowledge was available in computer-usable form about acoustic characteristics of spoken sentences or the regularities of English sound structure (phonological rules) that must be incorporated by a machine that handles continuous speech. Structural analysis of word sequences was based on typewritten text, and could not deal with errors or the fact that some spoken words are easier to identify than others. Little was known about what made one continuous speech recognition problem harder than another. Not much attention had been given to rapid processing of continuous speech, so that the few speech analysis algorithms that did work were extremely slow. The ARPA SUR project thus began in a context of only modest successes in continuous speech analysis and substantial uncertainty about whether versatile sentence understanding was going to be possible in the foreseeable future.

(4) An ad hoc study group of artificial intelligence contractors and speech researchers proposed the long range ARPA SUR PROJECT to explore the feasibility of a continuous speech recognizer that used linguistic and task-dictated constraints to aid the acoustic phonetic recognition procedures. Initially, the project involved five system contractors: Bolt Beranek and Newman, Inc. (BBN); Carnegie-Mellon University (CMU); Lincoln Laboratories of Massachusetts Institute of Technology; System Development Corporation (SDC); and Stanford Research Institute (SRI). After a mid-term evaluation of preliminary systems, BBN developed the "Hear What I Mean" (HWIM) system, CMU developed the Harpy and HEARSAY II systems, and SDC and SRI cooperated in the development of a system, only some components of which were implemented in the final SDC system. Supporting research efforts were also conducted at Haskins Laboratories, Speech Communications Research Laboratory, Sperry Univac, and the University of California at Berkeley.

The ambitious system specifications for this project, and the resulting performances of four final systems, are shown below:

GOAL	RESULTS WITH 1976 ARPA SUR SYSTEMS			
	HARPY	HEARSAY II	HWIM	SDC
Accept continuous speech, ...	184 sentences	22 sentences	124 sentences	54 sentences
from many cooperative speakers, ...	3 male, 2 female	1 male	3 male	1 male
in a quiet room, ...	(computer terminal room)			quiet room
with a good microphone, ...	(inexpensive close-talking mike)			good mike
with slight adjustments for each speaker, ...	20 training sentences	60 training sentences	no training	no training
accepting 1000 words, ...	1011	1011	1097	1000
using an artificial syntax, ...	BF=33	BF=33 or 46	BF=196	BF=105
yielding less than 10% semantic error, ...	5%	9% or 26%	56%	76%
in a few times real time (=300 MIPSS)	28 MIPSS	85 MIPSS	500 MIPSS	92 MIPSS

Harpy basically met or exceeded the system goals by achieving 95% correct understanding for the test set of 184 sentences from five talkers, using a 1011 word vocabulary and a highly-constrained syntax. The tests with around one hundred sentences do not provide a fully adequate performance evaluation, and 5 speakers is not exactly "many", but the use of a somewhat noisy room and an inexpensive microphone were beyond the system requirements. The syntax was heavily constrained, as indicated by the branching factor (BF) of 33, which is a language complexity measure based on the average number of words that could appear next in an acceptable sentence of the interactive language. Harpy required 28 million (computer) instructions per second of speech (MIPSS), which comes out to about 28 times real time on a large machine capable of 1 million instructions per second. It has subsequently been speeded up for rapid analysis and response on a minicomputer. Harpy contributed two primary ideas, including an "integrated network" for representing the expected pronunciations of total sentences, and a "beam search strategy" which allowed efficient testing of how close the incoming message corresponded with several similar sounding acceptable sentences.

"Semantic error" in a speech understanding system means misinterpreting some aspect of the meaning of the sentence, so that the wrong computer response would result. Insignificant errors in wording of the sentence do not then matter, provided the ultimate computer response is correct.

The other ARPA SUR systems (particularly the HWIM system) had higher error rates primarily because they dealt with more difficult tasks (larger branching factors) and were not as carefully tested and adjusted before the final demonstration tests.

On the whole, the successful attainment of the original ambitious system goals was a major contribution of the ARPA SUR project. Other primary accomplishments were in the development of several alternative system structures for coordinating knowledge from acoustics, phonetics, word matching, syntax, etc. While Harpy's integrated network and beam search strategy were important, and provide a "benchmark system" for comparatively evaluating future systems, the HEARSAY II and HWIM systems also were significant as moderately successful multiple-knowledge-source systems that are more readily extended to larger tasks than Harpy is. Many other systems were developed during the course of the project, but are now less relevant to future work. Expert respondents to our survey consider HARPY, HEARSAY II and HWIM (and the independently developed IBM system) to be the most important systems for future work.

Two other major areas of contribution were in (a) the study of how system performance related to linguistic constraints in the system, and (b) the development of important system <u>components</u> or knowledge sources. The branching factor as a measure of recognition task complexity was useful, though subsequent research is suggesting other even better measures of complexity. The correlation of system performance with branching factor was one indicator of the value of linguistic constraints in easing recognition. Also, the final results with the systems showed that even when accuracy of phonetic identification was quite low (e.g., 42% for Harpy) the accuracies of word identification (97%) and semantic

understanding of the total sentence (95%) were high. Structural constraints were thus useful in recovering from phonetic errors. Other studies showed that the HEARSAY I system, which could operate with or without syntax and semantics, worked significantly better when syntax and semantics were each introduced into the system. Of some interest was the evidence that system performance did not degrade much with doubling of the size of the vocabulary.

System components that were significant developments included:

- The "allophonic templates" or sub-phonemic segments that were used in Harpy to distinguish the alternative pronunciations of vowels and consonants in various contexts;
- The "phonetic lattice" used in HWIM to represent alternative segmentations and labelings of vowels and consonants that are consistent with the acoustic data in various regions of an utterance;
- The lexical decoding network in HWIM, which efficiently accounts for alternative pronunciations of words and the effects of each word on the pronunciations of neighboring words;
- Word verifiers in HEARSAY II, HWIM and SDC systems, which provide a second chance to compare the acoustic data with expected pronunciations of any hypothesized words; and
- Syntactic parsers that can determine the correct word sequence and structure of a sentence, even in the presence of errors in hypothesized wording, and with bidirectional analysis beginning at arbitrary positions throughout the sentence.

Other significant contributions from the ARPA SUR project were in the areas of system control strategies and experimental research. We recommend that the reader look at Figure 2-3 (page 19) of this report at this point, to see a listing of many of the contributions, and our assessment of their relative significances. A complex project of the size and interdisciplinary character of the ARPA SUR project necessarily produces a variety of contributions in various aspects of the problem. There is a real danger that Harpy's success and the usual attention given to recognition accuracy may overshadow many of the other primary contributions from the ARPA SUR project, but our relative assessments in Figure 2-3 should provide a more balanced picture. Our survey indicated that experts ranked the following areas among the primary contributions of the project (in decreasing order of significance):

- Control strategy and the integration of various knowledge sources;
- Segmentation and labeling of phonetic units;
- Word matching (and verification) procedures;
- Phonological rules;
- Prosodic analysis;
- Acoustic phonetic analysis; and
- Scoring procedures.

It is beyond the scope of this executive summary to explain these contributions and their relative significances, but the multiple forms of ARPA SUR contributions should be apparent. It is, however, useful to consider the following general advances in the state of speech recognition technology that were brought about by the project:

- At least one successful continuous speech recognizer (Harpy) now can handle vocabularies of 1000 words and accurately understand 95% of the sentences relevant to a modestly complex sentence understanding task;
- Several multiple knowledge source systems are available, with
  - Acoustics, phonetics, phonology, word matching, syntax, semantics, pragmatics, and some prosodics;
  - Clear evidence about the long-suspected utility of syntax and semantics;
  - Speech parsers that handle errors and arbitrary starting points in the structural analysis of the sentence;
  - Guidelines for prosodic aids to recognition;
  - Large lexicons and phonological rules that handle pronunciation variabilities and word boundary effects;
  - Improved detectors of various vowels and consonants, and a choice between allophonic templates or a phonetic lattice.
- More data have been compiled (and used) concerning the acoustic characteristics of sentences (especially regarding prosodics and some coarticulatory aspects of phonetics);
- Large lists of phonological rules have been compiled, tested, and used in recognition schemes;
- Useful measures of complexity and system performance have been developed; and
- Advanced scoring procedures have been developed that permit hypotheses from various aspects of the system to be uniformly scored and systematically selected, and strategies have been developed that assure the best possible interpretation of the phonetic structure of an utterance.

If it were possible to superimpose or set side-by-side the "before" and "after" pictures of the state of recognition technology in 1971 and 1976, it would be apparent how the ARPA SUR project contributed major advances (cf. the discussion on pages iv and vii of this summary). It is no wonder the experts we surveyed concluded that the project was important, needed, and productive of a significant advancement bordering on a "breakthrough". Yet in almost every component or aspect of a recognition system, there still is need for further major improvements.

(5) <u>CURRENT SPEECH RECOGNITION TECHNOLOGY</u> is the product of 26 years of research, development, and commercial initiatives, including several significant advances since the ARPA SUR project ended. Notable among the recent developments is the emergence of seven or eight commercial sources of limited speech recognizers:

- Heuristics, Incorporated offers a $299 hobbyist "SPEECHLAB", consisting of a hardware interface for small computers and suggested programs for word recognition;
- Phonics, Incorporated provides the cheapest stand-alone 16-word recognizer, for $550;
- Centigram, Incorporated offers a "MIKE" isolated word recognizer for $3000 to $5000, which interfaces to the ADAM computer;
- Interstate Electronics Incorporated markets a version of the former Scope Electronics "VDETS" recognizer for moderate size vocabularies, for around $20,000;
- Threshold Technology Incorporated, has for years led the industry with their VIP100 and TTI500 and TTI600 recognizers for highly accurate small vocabulary recognition at costs around $10,000 to $80,000;
- Dialog Systems, Incorporated offers several word recognizers which focus on use over telephone channels, and their Model 810 paging system sells for about $70,000;
- Nippon Electric Company has announced a limited "continuous speech recognizer" that handles both isolated words and restricted forms of connected speech such as digit strings, for a projected cost of around $80,000.

Perception Technology Incorporated is also said to offer recognizers, but we have no information about any systems being sold. These various recognizers are primarily intended for speaker-adaptive use, requiring a moderate amount of system training for each new talker. Costs of the more advanced systems are expected to come down dramatically in the next few years.

Word recognizers have been effectively used for various hands-busy applications such as:

COMMERCIAL:
- Package sortation systems;
- Inspection and quality control;
- Voice instructions to machine tools;
- Voice actuated wheelchairs;
- Hands-free control of hospital room environmental conditions;

MILITARY:
- Cartography;
- Voice authentication systems;
- Training skilled communicators like air traffic controllers; and
- Simulations of cockpit communications.

Current technology also includes these industrial development projects for advanced capabilities in speech recognition:

- Bell Laboratories: syntax-directed recognition of sequences of isolated words; several recognizers of digit strings, spelled-speech, and other limited forms of continuous speech;

- Carnegie-Mellon University: extensions and transportable mini-computer versions of Harpy;
- IBM: large project on continuous speech recognition using statistical techniques and laser patent texts as a task domain; applied to Harpy's task, this system attained 99% correct recognition for one talker in a quiet room;
- ITT Defense Communications Division: new entry into the technology, dealing with speaker independent recognition, telephone bandwidth, and noisy speech for isolated word recognition, word spotting, and low-cost hardware;
- Logicon: applying isolated word recognizers and a new connected word sequence recognizer to tasks of training air traffic controllers and simulations of recognitions of conning officer commands; latest system based on a Harpy-like mathematical machine;
- Nippon Electric Company: connected word sequence recognition, with a commercial product and further development work;
- Sperry Univac: linguistically-based connected word sequence systems and word spotting systems
- Texas Instruments: digit string recognition (using check digits to correct errors), for automatic speaker verification.

Other notable work is being done in France, Germany, Italy, and Japan. The commercial interest in speech recognition seems to be expanding rapidly, perhaps considerably more rapidly than current military activity in this field.

(6) GAPS REMAINING IN SPEECH RECOGNITION TECHNOLOGY, and their relative priorities, are listed in Figure 4-1 (page 40) of this report, and are worthy of the reader's careful scrutiny. Further work is needed in almost every aspect of recognition, despite the advances that the ARPA SUR project and other recent work have produced. Experts answering our survey considered that the following were among the top-priority aspects of recognition that need attention (listed in descending order of priority):

- Acoustic phonetic analysis;
- Prosodic cues to linguistic structures;
- Performance evaluation;
- Using linguistics to constrain ambiguities;
- System tuning on extensive data;
- Fast or near-real-time processing;
- Scoring procedures; and
- Phonological rules.

As is evident from Figure 4-1 and these opinions, the "front end" analysis routines which transform from acoustics to phonetics, phonology, words, and prosodics, are among the top priority components on which to focus. In addition, all systems need extensive testing and performance evaluation, which in turn requires research on comprehensive measures of task complexity and the specific causes of recognition errors, plus fast processing techniques. Comparative evaluations of alternative speech recognizers must

be undertaken, particularly for commercial isolated word recognizers, and the available practical recognizers need to be applied to various DOD applications. Extensive thought must be given to systematically bridging the large gap between practical isolated word recognizers and long range work on ambitious speech understanding systems. Part of that task involves human factor studies of the accuracy and other system performance characteristics that are needed in various practical applications. Another part involves defining appropriate "next steps" in advancing speech recognition capabilities. Programs are needed, but already underway, in digit string recognition and recognition of formatted word sequences. Speech understanding systems are, however, not getting the attention they deserve. Also required is consideration of "technology transfer", to translate results of such advanced development projects into improvements in practical recognizers. Finally, there is a continuing need, intensified and clarified by the ARPA SUR advancements, to conduct needed research on acoustic phonetic, phonological, and prosodic characteristics of spoken sentences.

(7) <u>RECOMMENDATIONS FOR ADVANCING SPEECH RECOGNITION</u> can be characterized as means for bridging the gaps in current technology. We recommend the undertaking of at least four types of speech recognition programs, including

1. Application studies with available commercial recognizers, such as determination of system requirements and performance characteristics for various DOD applications;
2. Programs to evaluate and advance current techniques, including;
   - ___ Comparative evaluations of alternative input modalities;
   - *___ Comparative evaluations of alternative speech recognizers;
   - *___ Human factor studies of practical speech input situations;
   - ___ Handling realistic channel conditions (microphones, telephone, noise, etc.)
   - ___ Extensions of recognizer capabilities without major system re-design.
3. Programs for developing advanced systems, including
   - *___ Evaluating the need for continuous speech;
   - ___ Developing digit string and word sequence recognizers;
   - ___ Harpy-like limited speech understanding systems, including:
     - * Further testing and performance
     - * Additional knowledge sources, such as prosodics; and
     - * Incremental compilation and automatic language acquisition, to allow new words, constructions, and tasks to be readily incorporated into the recognizer's capabilities.
   - *___ Development and continued refinement of <u>research systems</u>, with:
     - * Moderately restricted (HWIM-like) tasks;
     - * Independent (readily changeable) knowledge sources;
     - * Uniform scoring procedures;

* A lexical decoding network and phonological analysis procedures;
* Prosodic aids to recognition;
* Improved semantic and pragmatic constraints;

—— Task-independent (IBM-like) continuous speech recognizers; and

*—— *Methods for fast processing of extensive data.*

4. Research on necessary concepts and knowledge sources
   - *—— Acoustic phonetic analysis
   - *—— Prosodic aids to recognition
   - *—— Performance evaluations and task complexities
   - *—— Phonological rules;
   - *—— Linguistic constraints on ambiguities; and
   - *—— Scoring procedures for selecting hypotheses.

Those aspects marked with asterisks (*) are currently not being given sufficient (if any) attention and thus deserve early consideration. Further details are available in section 5.3 of this report.

(8) MECHANISMS FOR UNDERTAKING THE NEEDED WORK must also be considered. One possible way to bridge the gaps in current technology might be another large scale coordinated project. If another large scale speech understanding project were undertaken, the experts we surveyed would favor development of several alternative systems that address a spectrum of problem complexities, with one system directed at an easy problem, another at a moderately difficult task, and another at a quite difficult, challenging task. Other organizational features endorsed were the use of supporting research efforts conducted by specialist contractors, plus mid-term evaluations of the systems, extensive performance evaluation of the systems developed, and close interactions and frequent interchanges among contractors. The experts also recommended these system design characteristics:

- a moderate vocabulary of several hundred words or more;
- 10 to 100 speakers;
- three levels of system or language complexity;
- practical input through close talking microphones, telephones, or other communication channels of various qualities;
- systems adjustable to the speaker with only a few utterances;
- more substantial use of semantic and pragmatic constraints;
- near-real-time operation; and
- accuracy of 95-99%,
- to be achieved within a project period of three to five years.

The primary differences from the ARPA SUR project goals are: a series of progressively more difficult tasks; more attention to practical needs like realistic input channels, many speakers, high accuracy in real time; and no programmed demand for success on a fixed deadline.

While the gaps in current technology do not necessarily demand another large-scale speech understanding project, the need for careful coordination among the various intertwined aspects of recognition certainly favors some sort of coordinated program with close interactions and cooperation among various groups of researchers or developers. A large scale, multiple-contractor research program of cooperative and competitive developments of alternative

systems with common goals and extensive computer network facilities does foster such valuable interactions. Another mechanism for stimulating interactions and cooperation might be to encourage a group, like the recently-established voice interactions Technical Advisory Group (or "TAG") set up among United States government funders of speech research, to individually and/or collectively oversee further speech understanding research and development.

We also recommend the establishment of two or three "speech science centers" with speech and linguistic expertise, powerful computer facilities and network capabilities, and mechanisms for visiting researchers to use such facilities to advance their work on various aspects of recognition and to incorporate their advancements into the resident recognition systems. One center might study extensions of Harpy-like recognizers while another incorporates the best of the features of HWIM within a HEARSAY II like structure with independently operating knowledge sources. Such speech science centers could act as clearinghouses for useful speech databases, reports, research results, phonological rules, scoring procedures, recognition algorithms, and system structures, and could always offer working recognizers as research tools and testbeds for further advances. The centers could offer seminars and workshops for acoustic phonetic analysis, prosodics, phonological rules, interspeaker differences, etc. If affiliated with universities, they could provide the training and on-the-job experience appropriate for excellent new scholars in the field. Other speech research and development programs could be undertaken at the centers, such as in speech synthesis, speech transmission systems, clinical speech studies, and linguistic analysis. With stable funding from a variety of sources, and visiting scholars on sabbaticals or brief collaborative interactions, plus excellent facilities, such speech science centers could provide the concentrations of excellence needed to bridge the many gaps in current speech recognition technology.

Given the lead time needed to complete research projects and transfer the research into system, research should begin as soon as possible on improved acoustic phonetic analyses, phonological rules, prosodic analysis methods, and performance metrics and evaluation procedures. A coordinated program in the development of recognizers of several distinct capabilities could then be progressively undertaken. We also recommend that careful attention be given to transferring the on-going results from the recognition development projects and research into various practical DOD applications.

1. INTRODUCTION TO SPEECH RECOGNITION

This is an opportune time to assess the technology of automatic speech recognition and understanding, and to consider what "gaps" remain that require further work in the field. One reason for such an assessment is the successful completion of the Speech Understanding Research (SUR) project sponsored by the Advanced Research Projects Agency (ARPA) of the United States Department of Defense (DOD). While some limited attempts had been made to automatically determine the phonetic structure and wording of connected speech, the "ARPA SUR project" was designed to provide a "breakthrough" in the handling of spoken sentences, by the use of high-level linguistic information and task-dictated constraints on what could be said to the machine. It is useful to comprehend the full impact of the ARPA SUR project, to clarify its implications on future work, and to understand how it fits within the total current state of the art.

Another reason for assessment now is the growing industrial interest in speech recognition. The number of available commercial products has more than tripled within the last two years, and several large industrial groups are developing powerful new recognition systems. Researchers in various disciplines are also directing increased attention to this field.

Finally, the primary reason for this assessment of speech recognition technology is the increased governmental interest in defining what to do next, and concern with relating the research in this field to actual practical needs (particularly within the Department of Defense). In addition to our AFOSR, ONR, and ARPA-sponsored review of speech recognition technology, other indicators of governmental interest in this field include: (1) a recent workshop on voice input technology (Breaux, et al., 1978), at which six government agencies described 21 speech recognition projects; and (2) the establishment of a governmental Technical Advisory Group (TAG) concerned with voice technology.

We present in this report the following series of logically successive steps leading towards specific recommendations for further work to be funded in speech recognition:

1. Arguments favoring voice input to computers;
2. Explanations of how much can and should be said to a machine;
3. A brief history of early (pre-ARPA SUR) work on speech recognition;
4. A review of the contributions of the ARPA SUR project;
5. A summary of current technology;
6. Lists of gaps in current technology, and their relative significances;
7. Recommendations for filling those gaps and advancing speech recognition; and
8. Suggested mechanisms for undertaking the recommended work.

The suggestions we present are based not only on our own experience and study of the field, but also on our understanding of the consensus of opinions expressed to us by over 100 speech recognition experts. We conferred during this contract with experts at the following organizations active in speech recognition work:

dialect and sex comprised another (third) problem listed by the study group, and you may note from Figure 2-2 that the Harpy system used both male and

ARPA SUR CONTRACTORS	SYSTEM DEVELOPERS	COMMERCIAL SUPPLIERS
Bolt Beranek and Newman	Bell Laboratories	Threshold Technology
Carnegie-Mellon University	IBM Thomas R. Watson Research Center	Interstate Electronics
Haskins Laboratories	ITT Defense Communications Division	Dialog Systems
Lincoln Laboratory	Texas Instruments	(Also, conversations with representatives of Heuristics, Inc.; Phonics, Inc.; Perception Technology; Centigram; and Nippon Electric Company)
Sperry Univac		
Stanford Research Institute		
Systems Development Corporation		

In addition, we have discussed some of the issues represented herein with representatives of the following government agencies: ARPA; FAA/NAFEC; NADC; NASA Ames; NOSC; NTEC; ONR: AFOSR; RADC. We distributed a 30 page questionnaire to 160 individuals known to be knowledgeable in this field, and received 34 detailed replies. Some of the ideas included herein were presented at technical conferences and at an ONR-sponsored workshop for government funders of speech recognition.

In the remainder of Section 1, we introduce basic preliminaries that demonstrate why speech recognizers warrant further attention and support. The value of voice input is summarized (Section 1.1), the problems with voice input are noted (Section 1.2), some practical applications are mentioned (Section 1.3), the early history of speech recognition is very briefly summarized (Section 1.4), a spectrum of different types of recognizers is described (Section 1.5), and the distinction between "recognition" and the recent term "understanding" is illustrated for one system structure (Section 1.6). This section thus provides a cursory background for then considering the impact of the ARPA SUR project (Section 2), followed by a summary of current (1978) technology in speech recognition (Section 3), the remaining "gaps" (or issues and problem areas) in speech understanding technology (Section 4), and our recommendations for advancing speech recognition work (Section 5). Specific consideration is given to DOD applications for speech recognition, in Section 6. An extensive list of references were used in our study during this contract, only part of which is given in Section 7. Appendices on detailed ARPA SUR contributions (Appendix A) and a detailed survey of expert opinions about speech recognition (Appendix B) are included to provide further technical support for the conclusions presented throughout this report.

Readers who are familiar with speech recognition work may well skip or skim Sections 1 and 3. Sections 2, 4, 5 and 6 are the heart of our conclusions about ARPA SUR work, current needs, and future work.

1.1. The Value of Voice Input

First let us consider why you might want to talk to a computer. (For more details, see Lea, 1979b.) As shown in (Figure 1-1), voice input to computers offers many advantages. Since speech is the human's most natural modality of communication we have the potential of machines that more fully accommodate to the human user, rather than perpetuating the trend of our mechanical slaves actually enslaving us in unwanted diversions into learning key punching, typewriting, and complex programming methods. Any user needs little or no training to talk to a computer, other than learning how to constrain himself or herself to say only those things the machine can understand.

HEARSAY II used the most constrained grammars, with a finite state language,

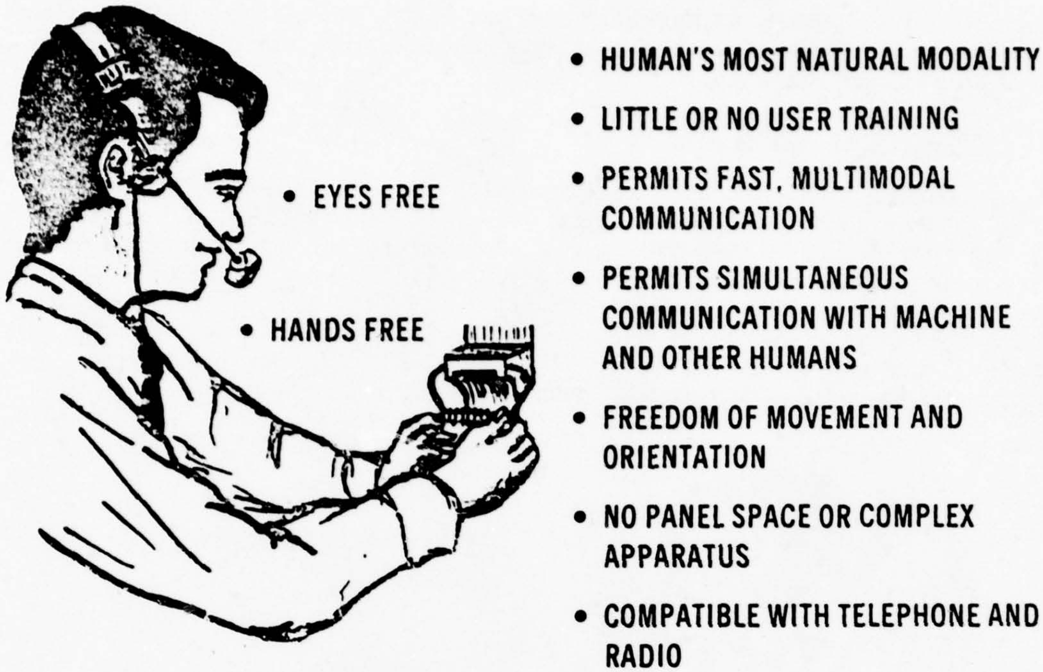


Figure 1-1. Advantages of voice input to machines.

Speech is our fastest communication modality. You can speak about twice as fast as the average typist can type, and human interaction experiments have shown that complex problems can be solved in half the time by teams of workers that communicate by voice, compared to other modalities (Chapanis, 1973; Chapanis, et al., 1977). Besides being our fastest modality, speech permits multimodal communication, so you can talk while pointing to a spot on a graphical input device or using buttons or other tactile devices. You can simultaneouly speak to machines and be understood by other humans. You have a freedom of movement and orientation that is unprecedented in computer input facilities. You can talk with your back to the machine and communicate at a distance, around obstacles, and in total darkness or blinding light. In some situations, such as for pilots in cramped cockpits, speech offers the distinct advantage of needing no instrument panel space or new devices other than a microphone. One of the most attractive prospects is the possibility of picking up a telephone and talking to a computer down the hall or across the country.

Perhaps the most important reason why over 800 speech recogniton devices have been sold in America, and why they have been used to process over 300 million words in factories and various field applications, is that they free the hands and eyes of the computer user. You can be busy measuring dimensions with a caliper, inspecting a product, handling packages, reading maps or instructions, monitoring assembly lines, controlling navigations, or otherwise engaged in the primary tasks of your job without being sidetracked to stop and input information to a computer.

constraints. The expected demand for large memory and computer power due

Voice input thus serves to markedly improve the productivity of the worker and relieve him or her from tedious interruptions that invariably lead to errors and unhappy workers.

1.2 Problems with Voice Input

Of course, there are some problems with voice input, also. As shown in Figure 1-2, commercial recognizers only work well with small vocabularies of a few tens of words like the digits or some special commands like mathematical operations. The larger the vocabulary, the more difficult the recognition task. Also, think of the difficulty you yourself can have in distinguishing some groups of words such as the rhyming letters of the alphabet, when spoken as "B, C, D, E, G, P," etc. Confusability is a function of the similarity in the sound structure of words.

- VOCABULARY (SIZE, CONFUSABILITY)
- TRANSDUCER AND CHANNEL CHARACTERISTICS
- SPEAKER VARIABILITY
  SEX, DIALECT, EXPERIENCE
- NOT PRIVATE
- ENVIRONMENTAL NOISE AND DISTORTIONS
- CURRENTLY COSTLY AND RESTRICTED

Figure 1-2. Problems with voice input.

Speech is also harder to understand when spoken over the telephone or other communication channels that restrict the bandwidth, add noise, and distort the frequency spectrum. What is even more important is the difficulty of understanding all speakers, whether they are male or female, from Boston, Harlem, Anaheim, or Dallas, and whether they are experienced in careful consistent articulations or not. The same talker may speak quite differently when he trains the machine to his or her voice, compared to later times when actually working on the job or when the voice changes due to a cold, physical or emotional stress, or even the time of day.

Speech is, of course, propagated indiscriminately, so it is not private, and computer security is difficult. Also, noises in the environment of the talker may be picked up, distorting the signal and making it difficult to identify. In the past, the best speech recognition facilities have not been cheap, compared to other computer input terminals. Yet, there are

The significance of the ARPA SUR project is partly reflected in the fact that HARPY, HEARSAY II, and HWIM were considered to be among the best current

definite signs that the costs are coming down fast, so that at least for the restricted forms of isolated word recognizers that you can buy today, the payoff from voice input facilities is really there. Also, initial costs for even the most expensive recognizers have been more than covered by reduced operating costs in practical human-to-computer interactions.

1.3 Early History of Speech Recognition

We can see that speech offers many advantages, but a few disadvantages that appear to be possible to overcome. Not surprisingly, the history of work in this field is almost as long as the history of electronic computers. As shown in Fifure 1-3, early work dates back to only a few years after the introduction of the well-known speech analysis tool, the sound spectrograph. In 1952, the first laboratory model of an automatic recognizer was developed, that identified which of the ten digits zero (spoken "oh") to nine was spoken by one specific speaker. That system used a two-dimensional templated based on the energies in two frequency bands above and below 900 Hz for each short time period, monitored throughout the utterance, as shown in Figure 1-3. The two-dimensional array of numbers for an input word was cross correlated (using analog electronic hardware) with stored training templates for each of the ten words, to decide to which one of the ten words it was most similar. Thus, the whole word was compared with training samples of the ten words. About 98% correct recognition was attained when trained for the speaker, but it dropped to 50-60% for a new talker. Several years later (Dudley and Balashek, 1958), a method was used to segment the speech into phonetic units, or time slices, like vowels and consonants, and slightly better recognition scores (over 99%) were reported, for the talker who trained the system. Early success in vowel and some consonant identifications was attained at Lincoln Laboratories (Forgie, 1959, 1961). The use of distinctive features was also attempted in the 1950's, providing a division of the two-dimensional pattern along the other dimension; namely, a division into simultaneous features (Wiren and Stubbs, 1956).

The first work using a digital computer came in 1960, along with the introduction of an inportant concept of time normalization, whereby short versions of an utterance that were spoken more rapidly than the training data were automatically stretched out or "normalized" to equal the normal duration of the training utterances, and slowly-spoken long versions could get reduced to a normalized length before comparisons and matching were attempted (Denes and Mathews, 1960). This recognizer also introduced a primitive form of "linguistic constraint" on expected pronunciation, based on "digram probabilities".

Through the following years, a large number of other recognizers were developed, as shown in Figure 1-3, (cf. also Martin, 1976; Reddy, 1976). Most were concerned with digit recognition, and many, such as the IBM "shoebox" recognizer, attempted to provide a compact working device that would demonstrate an initial capability in recognition. Indeed, throughout the 1960's, it was common for government funders of work in this field to be inundated by newcomers with "shoebox" or "suitcase" demonstration models of their entry into recognition work.

In 1965, a fairly successful scheme for recognizing and classifying vowels and consonants in continuous speech was developed (Hemdal and Hughes, 1965). An excellent review of the state of the art (Lindgren, 1965) called for higher-level linguistics to be used in recognition, so syntax could be used to guide a machine's choice of the wording of difficult utterances. This was not the first such call, but it popularized the concern for new sources of knowledge to aid recognition. The

* MET THE SYSTEM GOALS (CMU)........................ 12 11 10 9 8 7 6 5 4 3 2 1
· Accuracy (>90%)..............................

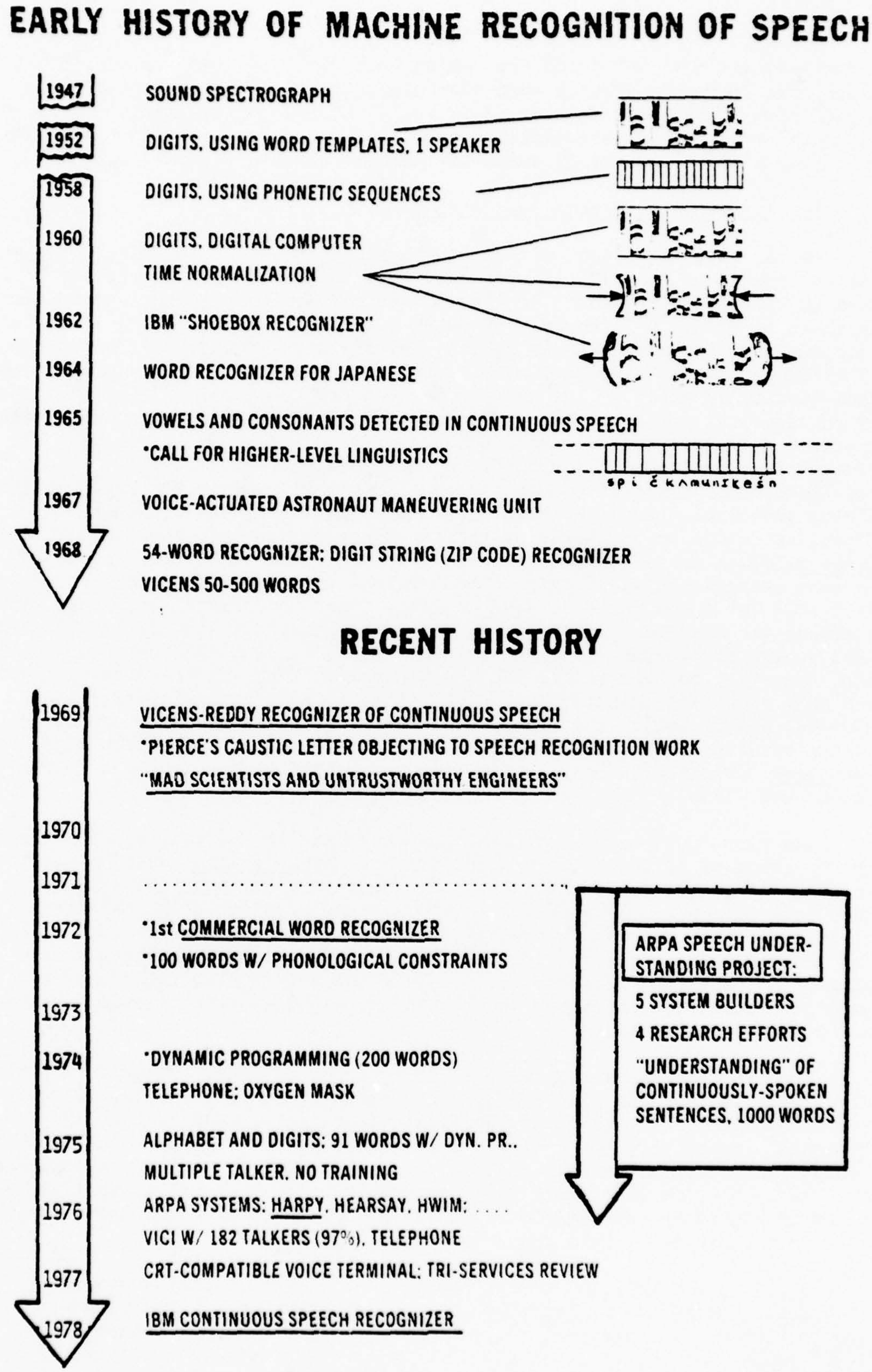


Figure 1-3. Some highlights in the history of speech recognition.

proposed applications for recognizers expanded to include a far-out idea of an astronaut manuevering unit (Kelley, et al., 1967) and zip code reading for the post office (cf. Martin, 1976). Expanded vocabulary word recognizers endeavored to handle vocabularies of 50 to 500 words (Bobrow and Klatt, 1968; Vicens, 1969), and achieved accuracies between 95 and 99% in the best of conditions.

In 1969, Reddy and Vicens introduced a recognizer of limited continuous speech (Vicens, 1969). Still the fledgling field was suffering growing pains and undertainties, and in 1969 the most popular letter to the editor that was ever published in the Journal of the Acoustical Society of America appeared. John Pierce (1969) of Bell Laboratories strongly objected to work in speech recognition and mused about its domination by "mad scientists and untrustworthy engineers." Pierce noted that not enough had been done to establish the real need for speech recognizers, and noted that repeatedly workers would undertake duplicative efforts without asking why they expect to succeed and why their work was worthwhile. "Re-inventing the wheel" or re-discovering old ideas seemed to characterize the field. For example, we may note the "single equivalent formant" (Focht, 1963), which was nearly equivalent to the dominant frequency components used in 1950 spectral anlyzers (and which seems to have been re-discovered but made more theoretically plausible in the "front cavity resonance"; Kuhn, 1975). Similarly, the old "vowel diagram" or "vowel triangle" of articulatory phonetics was manifested in the Dreyfus-Graf "sonograph" (1949), uses (e.g., Davis, et al., 1952) of plots of the lowest-frequency vocal tract resonance (formant 1) versus the frequency of the second lowest resonance (formant 2), and again in the vowel wheel of Yilmaz (1967).

Pierce also questioned the utility of small isolated word recognizers, and questioned how versatile continuous speech recognizers could ever work without extensive use of the human's sophisticated knowledge of language constraints and sense about the "meaning" of a conversation. Lea (1970) responded to Pierce's criticisms with several constructive suggestions for how to establish the value of speech input to machines and how to determine how much recognition ability is needed for specific applications.

One might have expected that the strong criticisms by such an influential researcher as Pierce could signal a setback in speech recognition work, especially since Pierce had earlier been instrumental in reducing government funding of the field of mechanical translation of languages (cf. Pratt, et al., 1966). In some ways the field did flounder, but within several years Bell Laboratories itself was (again) doing its own work in speech recognition. What's more, in 1971 the largest single project ever undertaken in speech recognition was begun, when the Advanced Research Projects Agency (ARPA) of the Department of Defense started a speech understanding research (SUR) project to develop machines that were capable of "understanding" continuously-spoken sentences involving a 1000-word vocabulary. That project was in a real sense out of context with most of the other work in speech recognition, as shown by the extra arrow in Figure 1-3. The focus was on bringing advances in artificial intelligence and computational linguistics to bear on the task of having the machine comprehend the full linguistic structure, and producing the intended machine response appropriate to the meaning of a sentence or discourse (Klatt, 1977).

The systems converged on the idea of probabilistic (or log likelihood)

While the ambitious ARPA SUR effort was going on, other important events occurred in recognition. The first commercial product appeared in 1972, and successful large-vocabulary word recognizers were built using rules of English sound structure or phonology (Medress, 1972), plus advances in dynamic programming (Itakura, 1975; White and Neely, 1975). One speaker-independent recognizer (called "VICI") correctly recognized 97% of the words spoken by any of 182 talkers (Scott, 1976). Recognition over the telephone was attempted, with some success (Itakura, 1975; Scott, 1976; Flanagan, et al., 1979). In 1977, a voice terminal that looks to a computer just like a standard CRT terminal, was offered commercially by Threshold Technology. Throughout much of the history of speech recognition, IBM has been involved in a series of valuable projects, and at this writing in 1978, they have the largest ongoing project in speech recognition (Bahl, et al., 1978).

1.4 Types of Recognizers

The result from this 26 year history is a variety of different recognition capabilities, as summarized in Figure 1-4. The easiest task is recognition of isolated words, surrounded by pauses, and then one can use linguistic constraints on allowable sequences of such words, by still maintaining pauses, as in the sequence "...RIGHT...THIRTY...DEGREES..." When the speech flows freely in connected form, word boundaries are hard to find and words distort the pronunciation of the neighboring words, but the task can be limited by handling only strings of digits ("...ZERO THREE ZERO...") or strictly formatted word sequences ("...RIGHT 30 DEGREES..."), where only certain words can go in each slot in the structure. Key information-carrying words can be spotted in the context of conversations, so that one can decide that a conversation is worth listening to based on the occurrence of certain key words like "NIXON" and "WATERGATE". Restricted speech understanding is intended to handle total sentences relevant to a specific task, using all the knowledge that can be brought to bear about the grammatical and plausible sentences one can say when working in a specific task domain. Task-independent continuous speech recognition will allow a change in topic from one utterance to the next and still correctly identify the wording of the sentences, such as illustrated in Figure 1-4 for the sentences "What is the big event in Anaheim?" and "How many aircraft carriers does Russia have?

- **ISOLATED WORDS** ...(Pause)...RIGHT...(Pause)
- **SEQUENCES OF ISOLATED WORDS, USING LINGUISTIC CONSTRAINTS** ...RIGHT...THIRTY...DEGREES
- **CONNECTED SEQUENCES OF**
  - —DIGITS, OR ...ZERO THREE ZERO...
  - —WORDS IN STRICT FORMATS ...RIGHT 30 DEGREES...
- **SPOTTING KEY WORDS IN CONTEXT** xxxxNIXONxxxxxxWATERGATExxxx
- **RESTRICTED SPEECH UNDERSTANDING** "Tell me about Nixon and Watergate."
- **TASK-INDEPENDENT CONTINUOUS SPEECH**

"What is the big event in Anaheim?"... "How many aircraft carriers does Russia have?

Figure 1-4. Forms of speech recognized by machines.

1.5 Recognition vs. Understanding

Figure 1-5 shows the basic distinction between "recognition" and "understanding" systems, and also illustrates the basic building blocks involved in machine analysis of speech, based on a conventional structure called a "bottom up" recognition scheme. The acoustic speech signal is first analyzed to extract basic acoustic parameters like the frequency spectrum and the energy in different time segments. Then information carrying features are extracted that define various phonetic events like the positions of various vowel-like sounds, whether or not the talker's vocal cords are vibrating, and how noisy or fricative-like the speech signal is. Then that information is used to divide the speech into time slices or segments, which are labelled with phonetic categories, like "this is an s-like sound" and "the next is an ee-like sound," and so forth. The phonetic sequence for the input speech is matched to stored sequences of expected pronunciations for each of the words in the lexicon or dictionary,

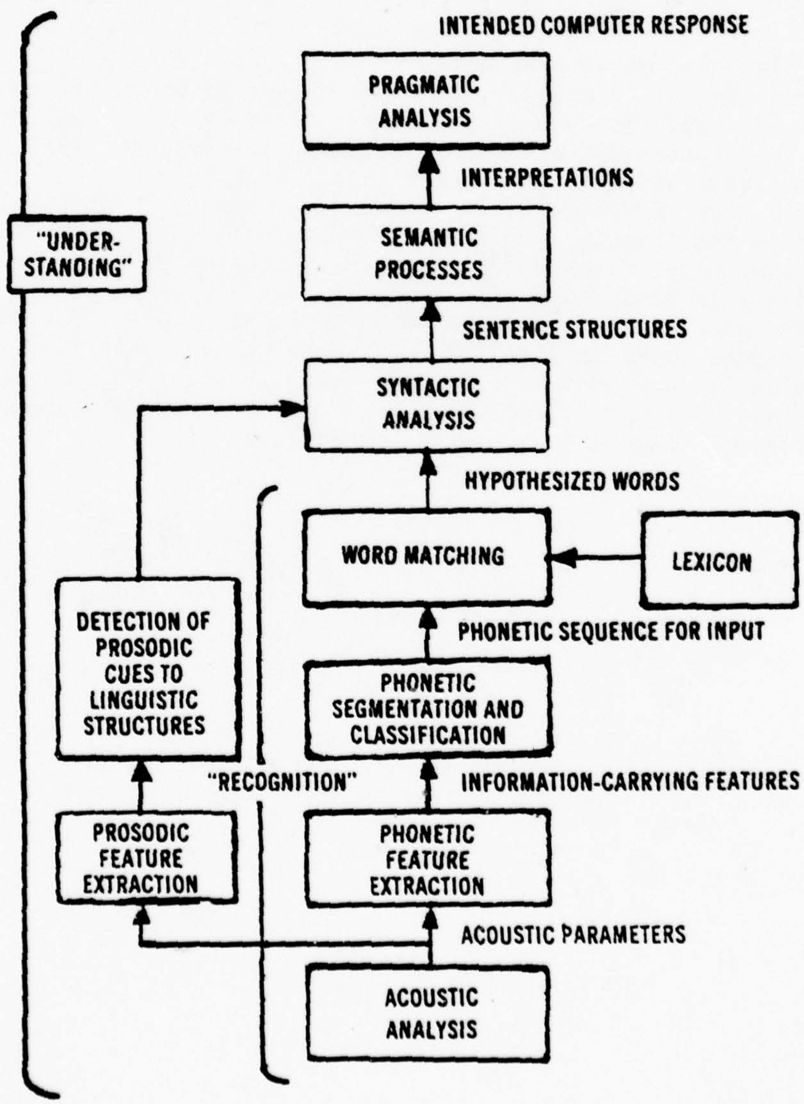


Figure 1-5. Models of "recognition" and "understanding"

and the best matching sequences are considered the most probable words to have occurred in the speech. Even in the earliest work in recognition, such hypothesization of words was expected to be followed by some determination of sentence structure and meaning, but the designers of the ARPA SUR project coined a new phrase, "speech understanding," for the more complete system which uses syntactic analysis, semantics, and pragmatic information to finally lead to an appropriate computer response, such as the performing of a calculation or the retrieval of data from a stored database.

Speech understanding systems were also expected to use prosodic features like the pitch or intonation of the talker's voice, the stress or emphasis placed on various syllables, and the rhythm and rate of the speech. Such information was expected to be useful for detecting large-unit linguistic structures like boundaries between phrases, differences between commands and questions, and special grammatical structures like subordination and conjunction of phrases.

All of these knowledge sources may be integrated into <u>various</u> system structures and control strategies, so that guesses based on incomplete knowledge in one area can be reinforced or altered by the other forms of knowledge. The "bottom-up" hierarchy illustrated in Figure 1-5 is only one of many alternative system structures. Others will be discussed in Section 2 and Appendix B, for example. A major concept of speech "understanding" systems is that the incoming acoustic information is not enough, and several knowledge sources must cooperatively select hypotheses so that the ultimate correct <u>response</u> is obtained from the system, even if it makes mistakes along the way, gets misdirected by a wrong hypothesis, or misidentifies some insignificant words in its final decisions about the wording of the sentence. Thus, it can be seen from Figure 1-5 that the bottom four boxes, in general, represent the task of speech recognition, and the entire group of boxes represent that of speech understanding.

2. SPECIFIC CONTRIBUTIONS OF THE ARPA SUR PROJECT

2.1 Introduction

We focus here on the specific contributions of the largest project ever undertaken in speech recognition; namely, the nine-contractor Speech Understanding Research project sponsored by the Advanced Research Projects Agency of the United States Department of Defense (the "ARPA SUR project"). The detailed designs and performances of the four final recognition systems are presented in previous literature and within a forthcoming book (Lea, 1979a), in Chapters 12-16, and an overview of the project appears in Chapter 11 of that book. What we add here is a detailed assessment that provides the reader with lists of ideas, techniques, results, and comparisons which can be relevant in future work. In this section we focus on generalities, which are discussed further in Appendix A. Thus, Section 2.2 gives a brief overview of the limited capabilities in recognition that formed the context in which the project began, and a summary of how the major goals defined for the project were met or exceeded. In Appendix A we list some detailed contributions in system design (Section A-1), and detailed contributions to various components or aspects of recognition (Section A-2). Readers interested only in general contributions might well skip or skim over Appendix A. We venture in Section 2.3 our list of the primary contributions of the ARPA SUR project. The expected impact on future work is summarized in Section 2.4.

2.2 Before and After ARPA SUR

During two decades of research prior to the ARPA SUR project, there had been repeated calls for overcoming the major hurdle separating moderately successful isolated-word-recognition systems from the unattained ideal of more natural uninterrupted voice communication with computers. Review articles had repeatedly called for the full use of language structures such as acoustic phonetics, coarticulation regularities, phonological rules, prosodic structures, and especially syntax and semantics (Lindgren, 1965; Hill, 1971; Lea, 1972). The ARPA SUR project was the first large-scale effort to use artificial intelligence ideas and higher-level linguistic information to provide a technology for understanding spoken sentences. In this section, we take a look back at the context in which the project began, and assess how the initial goals were met.

2.2.1 The State of the Art in 1971

Figure 2.1 summarizes some important aspects of the state of speech recognition technology when the ARPA SUR project began in 1971. No one had shown that continuous speech could be recognized. Existing examples at that time (e.g., Vicens, 1969) were too limited to be extendable. Isolated word recognition had become quite accurate (over 95% correct) in a few prototype systems using high-quality speech, but their techniques had not been employed to reliably recognize words in sentence contexts. Nothing like a 1000-word vocabulary had ever been attacked, and it seemed that such large vocabularies could lead to combinatoric problems of major magnitude. No complete systems had been built using multiple forms of incomplete knowledge like acoustics, phonetics, lexical processing (word matching), prosodics, syntax, semantics, and pragmatics, plus appropriate control structures. Only the acoustic phonetics and word matching procedures were actually functional in available recognizers.

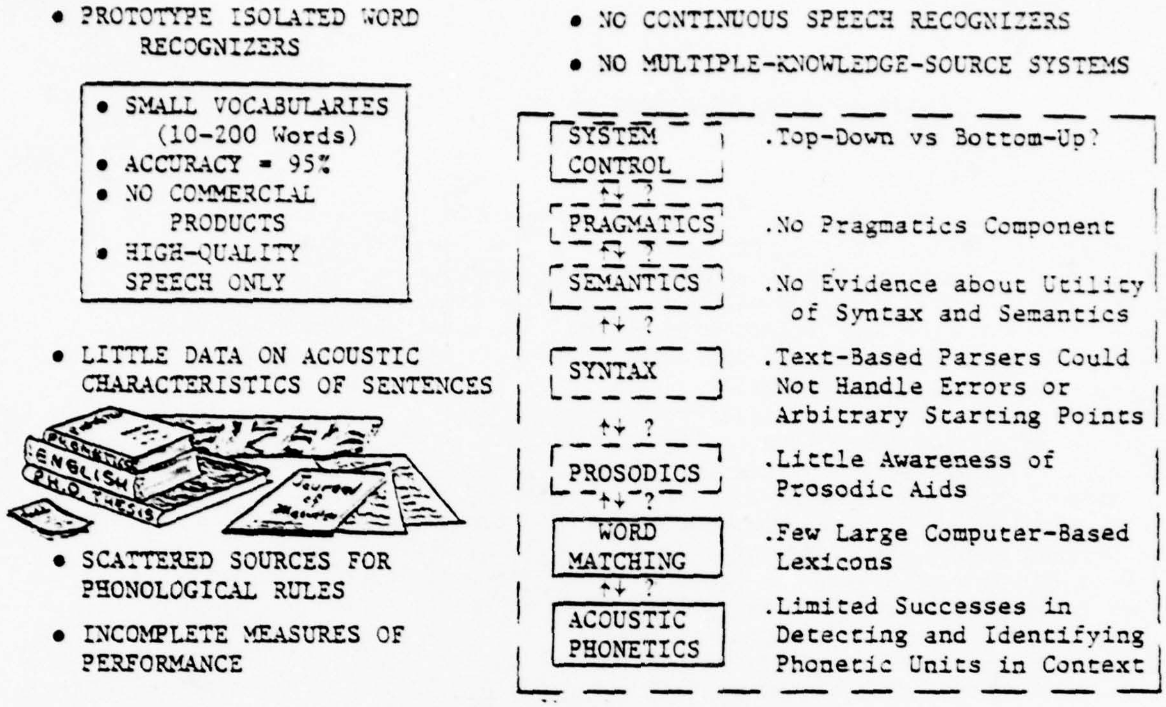


Figure 2-1. The state of ASR at the beginning of the ARPA-SUR project (1971)

Not only were there no adequate systems in 1971; even the necessary knowledge, ideas, and experimental data were limited and spotty. Very little data were available on the acoustic characteristics of phonetic segments or prosodics in spoken sentences. Scattered throughout the linguistic literature and various acoustic phonetic studies were phonological rules that had not been compiled in one place or used in speech analysis or recognition. There was also considerable uncertainty about how to use phonetic and prosodic data and phonological rules along with higher-level linguistic constraints. The only syntactic parsers were text-based, and they could not deal with errorful input strings or parsing from arbitrary starting points in the sentence. "Top-down" control strategies (whereby syntax first hypothesizes words that are then verified from acoustic data) were being suggested as a promising alternative to traditional bottom-up strategies (which did phonetic analysis first, hypothesized words, and finally would weed out wrong word sequences with linguistic constraints). Other system strategies were possible, but untried.

While researchers expected final accuracy of semantic understanding to be different from accuracy of phonetic segmentation and word matching, they did not know just how much improvement syntax and semantics would make, or how low phonetic classification accuracy could be and still allow adequate accuracy of semantic understanding. They had little more than the vocabulary size and number of words in a sentence that could help gauge the complexity of a recognition task, so they didn't know whether 70% accuracy on a hard task was better or worse than 90% accuracy on an easier task. Most recognition techniques were not real-time processes, and the speed of complex speech understanding systems could only be guessed at.

The ARPA SUR project brought some answers in each of these problem areas, and others as well. This is not to say that other projects have not also made major contributions since 1971, or that there were not others also aware of the problems addressed by the ARPA SUR project. Indeed there were, but general review of other work is not our concern here. (See Lea, 1979 c, for a review of other relevant work.)

2.2.2 Meeting the Original Goals

In the spring of 1970, an adhoc study group was formed in response to a request by Dr. Larry Roberts, then Director of the Information Processing Techniques Office (IPTO) of the Advanced Research Projects Agency. The study group consisted of experts taken from the then-current ARPA contractors working on artificial intelligence, and consisted of Allen Newell and Raj Reddy of Carnegie Mellon University, James Forgie, Dennis Klatt, and J.C.R. Licklider of MIT, Jeffrey Barnett of Systems Development Corporation, John Munson of Stanford Research Institute, William Woods of Bolt Beranek and Newman, and Cordell Green of ARPA IPTO. The committee was charged with exploring the "feasibility of demonstrating a speech recognition system with useful capabilities and greater power than current isolated word recognition systems". They concluded that a reasonable chance to achieve such a system would take at least five years, requiring: (1) major technical advances in the systematization and use of acoustic-phonetic and phonological structures; (2) cooperative efforts by several technical disciplines (including acoustics, phonetics, phonology, syntax, semantics, and task constraints); (3) intermediate experimental systems to be demonstrated around the mid-term of the project; and (4) a set of ambitious but reasonable specifications for evaluating the performance of the final systems. The study group, and the project itself as first initiated, did not require that the project provide a careful investigation of the potential uses of speech understanding in practical DOD applications, but the study group did expect that the demonstration of technical feasibility would represent a significant step toward a capability suitable for practical applications (Newell, et al., 1971, p. 3).

Given the fledgling state of continuous speech recognition in 1971, and the defensive posture the field had following Pierce's (1969) pessimistic evaluation of speech recognition work, the goals defined by the ARPA study group were very ambitious. Yet, the study group recognized that some practical constraints had to be placed on the overall task of understanding continuous speech if useful systems were to emerge in the near future. They thus defined the goals shown in Figure 2-2, and called for their fulfillment in a five year intensive program.

Occasionally we have heard the ARPA SUR project criticized in terms of how it missed a target defined post facto by those who wished the project had attacked problems it did not set out to deal with. We think it is first appropriate to consider how the project results related to the initial goals, and how the difficulties encountered compared with the initial projections of the study group. The study group outlined 19 technical problems (or "dimensions of difficulty") involved in recognition, some of which are shown in Figure 2-2, along with results for the final four systems.

2.2.2.1 Continuous Speech - One of the primary "breakthroughs" sought was to show the feasibility of understanding continuous speech, and each of the final systems was demonstrated (with varying degrees of accuracy) by handling

GOAL : ACCEPT CONTINUOUS SPEECH FROM MANY COOPERATIVE SPEAKERS,

HARPY : Primary test with 184	sentences from	3 Male, 2 Female	speakers,
HEARSAY II: Tested with 22		1 Male	
HWIM : Tested with 124		3 Male	
SDC : Tested with 54		1 Male	

GOAL : IN A QUIET ROOM, WITH A GOOD MIKE, AND SLIGHT TUNING/SPEAKER,

HARPY :	in a computer terminal room, with a close talking mike, and	20	training sentences per speaker,
HEARSAY II:		60	
HWIM :		NO	
SDC :	in a quiet room, with a good mike, and	NO	

GOAL : ACCEPTING 1000 WORDS, USING AN ARTIFICIAL SYNTAX & CONSTRAINING TASK,

HARPY :	1011 words, finite state language,	BF=33	for document retrieval,
HEARSAY II:		BF=33,46	
HWIM :	1097 words, restricted ATN grammar,	BF=196,	for travel management,
SDC :	1000 words, context-free grammar,	BF=105,	for data retrieval,

GOAL : YIELDING < 10% SEMANTIC ERROR, IN A FEW TIMES REAL TIME (=300 MIPS)

HARPY :	yielding	5%	semantic error, with	28 MIPS
HEARSAY II:		9%, 26%		85 MIPS
HWIM :		56%		500 MIPS (1350 R.T. on .35 MIPS)
SDC :		76%		92 MIPS

Figure 2-2. Goals and performance characteristics for final (1976) ARPA SUR Systems.

a modest number of spoken sentences. There is little or no doubt that restricted forms of continuous speech recognition are now possible. (The restrictions will be clarified as we discuss the other problems or dimensions listed in Figure 2-2.) Indeed, Harpy's success with continuous speech has encouraged the development of a Harpy-like integrated network for recognizing connected speech in computer-assisted training of air traffic controllers at the Naval Training Equipment Center and Logicon. The Harpy continuous speech task was later used as a benchmark task for evaluating the IBM continuous speech recognizer (Bahl, et al., 1978). All of the ARPA SUR systems, even including the intermediate systems like the Lincoln Laboratories system, the SRI system, SPEECHLIS, HEARSAY I, and the SDC vocal data management system, did directly attack the continuous speech problem, and none shirked from that responsibility. How well they did is another issue.

2.2.2.2 Muliple Speakers, Single Dialect - A second problem area addressed by designers of ARPA SUR was the multiple-speaker problem. The ARPA SUR systems were targeted to work with "many" speakers. Harpy was tested with 5 speakers, on its document retrieval task, and with 20 speakers on a 3-digit-string recognition task. HWIM dealt with 3 speakers in its final test, and incorporated a procedure for estimating the talker's vocal tract length for automatic talker normalization. While one to five speakers may not be "many" the problem of handling multiple speakers was directly addressed, with considerable success, at least for Harpy. In contrast, after 15 years of various projects, IBM and other long-term contributors to the field were still handling the speech of only one male speaker in 1978 (Bahl, et al., 1978). Speaker

dialect and sex comprised another (third) problem listed by the study group, and you may note from Figure 2-2 that the Harpy system used both male and female talkers of a similar dialect, as originally called for in the project design. The limited populations of speakers shown in Figure 2-2 are more an indication of the limited testing of the systems than they are an indication of inability or unwillingness to handle multiple speakers.

Many applications, such as military systems and even many factory installations, permit or require only a few talkers to speak to the machine, so that systems (like the 1976 ARPA SUR systems) which can handle only one or a few talkers are quite acceptable. Indeed, most commercial recognizers are speaker-dependent, and some researchers (e.g. Neuburg, 1975) suggest that the multiple-speaker problem should not be given high priority, especially if a system can be readily tuned to a new talker.

2.2.2.3 Tuneability - A related design problem concerns how much training of the machine to the individual characteristics of the speaker is needed. Harpy and HEARSAY II need only about 20 to 60 sentences of training data before they can accurately recognize a new talker of similar dialect. HWIM and the SDC system actually tried to get along without any training to the individual talker. All these results were directly in accord with the original ARPA SUR goal of only "slight tuning to the speaker", and may be contrasted to systems like IBM's that require over one hour of training utterances to determine a priori probabilities and other required statistics for each speaker.

2.2.2.4 Input Environment - Environmental noise was another problem listed in the ARPA SUR plan, and on this aspect the final systems exceeded the initial goals, by dealing with somewhat noisy (65dBA) speech recorded in computer terminal rooms, rather than in antiseptic conditions of acoustically-insulated quiet rooms. Most laboratory models of recognizers (e.g., IBM's recent system; Bahl, et al., 1978) have been tested only with high quality speech.

While the original goals also called for easing the recognition difficulty by use of high quality microphones, Harpy, HEARSAY II and HWIM were tested with inexpensive close-talking, noise-cancelling microphones. Harpy was also tested with telephone speech. The telephone and microphone conditions are thus among those goals in the initial study that were dealt with by the best-performing of the ARPA SUR systems. Yet, further studies are needed to determine how system performances degrade with various amounts of input distortion such as noise, bandwidth, and mis-shaping of the frequency spectrum.

2.2.2.5 Vocabulary and Language Constraints - The project produced significant advances on other dimensions of the recognition problem, also. The systems were supposed to handle a large vocabulary of 1000 words, and they each did so. The 1000-word vocabulary was a major advance from the 16-word vocabulary used in previous continuous speech recognition (Vicens, 1969), or the small vocabularies (ranging only in a few cases above 100 words) used in speaker-dependent isolated word recognition. Perhaps equally important were the studies of how vocabulary size and other complexities of the task affect recognition performance, as is discussed further in Appendix A, Section A-3.3.

The original project plan also focused on how to use syntactic, semantic, and pragmatic information to support recognition. Systems were expected to use an artificial syntax which highly restricts what can be said. Harpy and

mechanical slaves actually enslaving us in unwanted diversions into learning key punching, typewriting, and complex programming methods. Any user needs little or no training to talk to a computer, other than learning how to constrain himself or herself to say only those things the machine can understand.

HEARSAY II used the most constrained grammars, with a finite state language, or Markov model. HWIM used a powerful grammar called an augmented transition network or ATN grammar, and the SDC system used a moderately difficult "context free" grammar. One way to measure the complexity of the recognition problem is the so-called branching factor, or "BF" shown in Figure 2-2, which is the average number of words that could appear next in an allowable sentence of the language. Thus, the larger the BF, the more difficult the task, though this one metric is hardly adequate as a full measure of complexity of a recognition task. Still, one reason for the success of the HARPY system, in particular, was its effective use of syntax to constrain the possible wordings that might reasonably be hypothesized and distinguished. In evaluating the limited successes of the HWIM and SDC systems, the complexities of their tasks must be considered. These questions of task and language complexity, and effective use of constraints, are discussed further in Section A-3.3 of Appendix A.

Semantic support, and the use of user models and modelling of the total discourse of interaction, had been called for in the initial study report, but they had less evident effects on the performance of the final ARPA SUR systems than did syntactic constraints. HEARSAY I, developed early in the project, had very strong constraints built in by the semantic and pragmatic information and the user model, and those constraints really helped recognition. SRI developed sophisticated semantics and discourse models that unfortunately were not incorporated into the final SDC system (Walker, et al., 1976; Walker, 1979).

2.2.2.6 Accuracy and Other System Performance Measures - Harpy exceeded the primary accuracy goal of the ARPA SUR project, by successfully understanding 95% of the sentences spoken to it, or, as is shown in Figure 2-2, making a semantic error 5% of the time. The HEARSAY II system also matched the accuracy goal for that limited task, but when tried on another more challenging task with a branching factor of 46, its error rate tripled, to 26%. The HWIM system, with its challenging task represented by the large branching factor of 196, was wrong in its interpretation of a sentence in slightly over half the sentences it was tested with. System Development Corporation lost its major computer system only months before the final demonstration, so they had to drop their plans to use higher-level linguistic components designed at Stanford Research Institute, and developed a substitute system that only attained 24% correct understanding.

Accuracy is the one system parameter that has dominated the evaluative discussion of ARPA SUR systems, and stimulated the extensive interest in Harpy. However, it is important to place accuracy results in the full context of task complexity, extendibility of the system to new tasks, and other goals, such as speed. The reader should refer to Chapter 14 (Section 14.4.4) and Chapter 15 of (Lea, 1979a) for some qualifications on Harpy's success.

While the goal was that the systems would give results in a few times real time on very large machines handling 100 million instructions per second (MIPS), it took minutes for responses to a few-second sentence in the fastest systems (working on moderate-size computers of less than 1 MIPS), and an hour or more for the extremely slow HWIM system. However, even the 1350 times real time required for HWIM, using an 0.35 MIPS PDP-10 computer, would correspond to only 5 times real time on the target 100 MIPS machines. Even though the computer technology had not advanced to providing the projected 100 MIPS, the systems were able to function under effective use of linguistic

job without being sidetracked to stop and input information to a computer.

constraints. The expected demand for large memory and computer power due to combinatorial explosions of hypothesized words did not materialize to the degree feared. For Harpy, the speed goal was exceeded by an order of magnitude (cf. Lowerre and Reddy, 1979). Harpy has subsequently been speeded up, so it now operates in near real time on a minicomputer with paged memory. Near-real-time analysis has been found to be useful not only for demonstrating the feasibility of practical systems, but also for speeding up research and permitting the processing of extensive data so that systems can be carefully tuned.

Not too many long-term projects succeed in matching or exceeding so many of their original goals, such as the ARPA SUR project did. Still, one could have hoped for a more thorough final performance evaluation than the hundred or so sentences with which even the most complete system tests were made. Such scientifically adequate evaluations were proposed but not funded.

2.3 Primary Contributions

In this section, we assess the various contributions of the ARPA SUR project that have been detailed in Appendix A. We consider some expert opinions about the ARPA SUR contributions (Section 2.3.1), give our assessment of the relative significance of the various contributions (Section 2.3.2), and provide a summary of the resulting advances in the state of the art (Section 2.3.3).

2.3.1 Expert Opinions About ARPA SUR Work

To guide us in our assessment of primary ARPA SUR contributions, we visited all ARPA SUR contractors and almost all of the active speech recognition groups in the United States, and conferred with other colleagues, including workers from Japan, France, Germany, Canada, Poland, and Australia. We interviewed over 100 workers in speech recognition, and distributed a 30-page questionaire to over 160 such workers (with 34 formal replies), soliciting opinions about the ARPA SUR project, the best current techniques in recognition, and the future needs and trends. Survey results are detailed in Appendix B. Some of the issues about which there was most agreement were that the ARPA SUR project was ambitious and needed, and that it resulted in a significant advancement bordering on a breakthrough.

When given a list of topics to rank order, the 34 respondents to the questionaire gave average rankings which indicated that primary contributions of the project (in decreasing order of significance) were in:

1. Control strategy and the integration of various knowledge sources;
2. Segmentation and labeling of phonetic units;
3. Word matching (and verification) procedures;
4. Phonological rules;
5. Prosodic analysis;
6. Acoustic phonetic analysis; and
7. Scoring procedures.

When the respondents were given the opportunity to write their own lists of primary contributions, the most frequently mentioned contributions could be grouped into the following categories (with numbers of explicit mentions as indicated): system structure and control (21 mentions); phonological rules (8); acoustic phonetics (7); parsing and higher-level linguistics (5) and prosodics (4).

The significance of the ARPA SUR project is partly reflected in the fact that HARPY, HEARSAY II, and HWIM were considered to be among the best current speech understanding systems, along with the independently-developed IBM system. A large majority agreed that the independent supporting research efforts conducted by specialist contractors constituted a good aspect of the project.

Other details of those expert opinions are beyond the scope of this overview, and there is some doubt about the representativeness of the 34 completed questionnaires, almost half of which came from former ARPA SUR workers. However, it is noteworthy that the respondents had an average experience of 10 years work in speech recognition, so they do represent an elite group of experts in the field. We instead present next our own overall assessments, for which we take sole responsibility, but which we believe generally reflect the opinions and evaluations of most of the 100 or more experts with whom we have conferred.

2.3.2 Relative Significances of Various Contributions

Figure 2-3 illustrates our attempt at assessing the relative significances of various contributions from the project. Contributions with scores of "1" are considered the most significant, then come those whose bars reach level "2", then "3", and so on. Notice that major categories of significance include: meeting the system specifications (Section 2.3.2.1), providing several system control structures for speech understanding (Section 2.3.2.2), studying system control and search techniques (Section 2.3.2.3), studying linguistic constraints and system performances (Section 2.3.2.4), developing needed components or knowledge sources (Section 2.3.2.5), and conducting experimental research about speech analysis techniques (Section 2.3.2.6). Our listing here of primary contributions is unavoidably redundant for those who have carefully studied the contributions listed in Appendix A, but the assignment of significances should be helpful.

2.3.2.1 Meeting the System Goals - It is certainly significant that the project met the challenging system goals for understanding continuously-spoken sentences. While the attained accuracy was a significant success, the handling of large vocabularies and dealing with all the peculiar coarticulations and ambiguities of continuous speech were also important. Researchers and system developers would be ill-advised to forget that this was accomplished by effective use of linguistic constraints. Of lesser, but still fairly high, significance were the moderate speed, practical input environment, and multiple speakers involved in the final system performances.

2.3.2.2 System Structures - Ultimately more significant than the accuracy of the final systems was the major stride in developing alternative system structures for speech understanding. The project focused attention on the large combinatorial space of alternative system designs, and showed advantages and disadvantages of many experimental system structures.

Harpy's success makes it an outstanding contribution, yet long after its 1976 accuracy scores are out of date, its value will linger in its methods of integrating knowledge into a composite pronunciation network and efficiently searching the network for acceptable pronunciations with the "beam search technique". It is the best available system for small recognizers of spoken sentences, and it will remain a benchmark system for assessing future systems.

Harpy has its drawbacks, including the complex and time-consuming processes of determining allophonic templates, determining all dictionary pronunciations, handling word-juncture phenomena, compiling the network, and

vowels and consonants in continuous speech was developed (Hemdal and Hughes, 1965). An excellent review of the state of the art (Lindgren, 1965) called for higher-level linguistics to be used in recognition, so syntax could be used to guide a machine's choice of the wording of difficult utterances. This was not the first such call, but it popularized the concern for new sources of knowledge to aid recognition. The

5

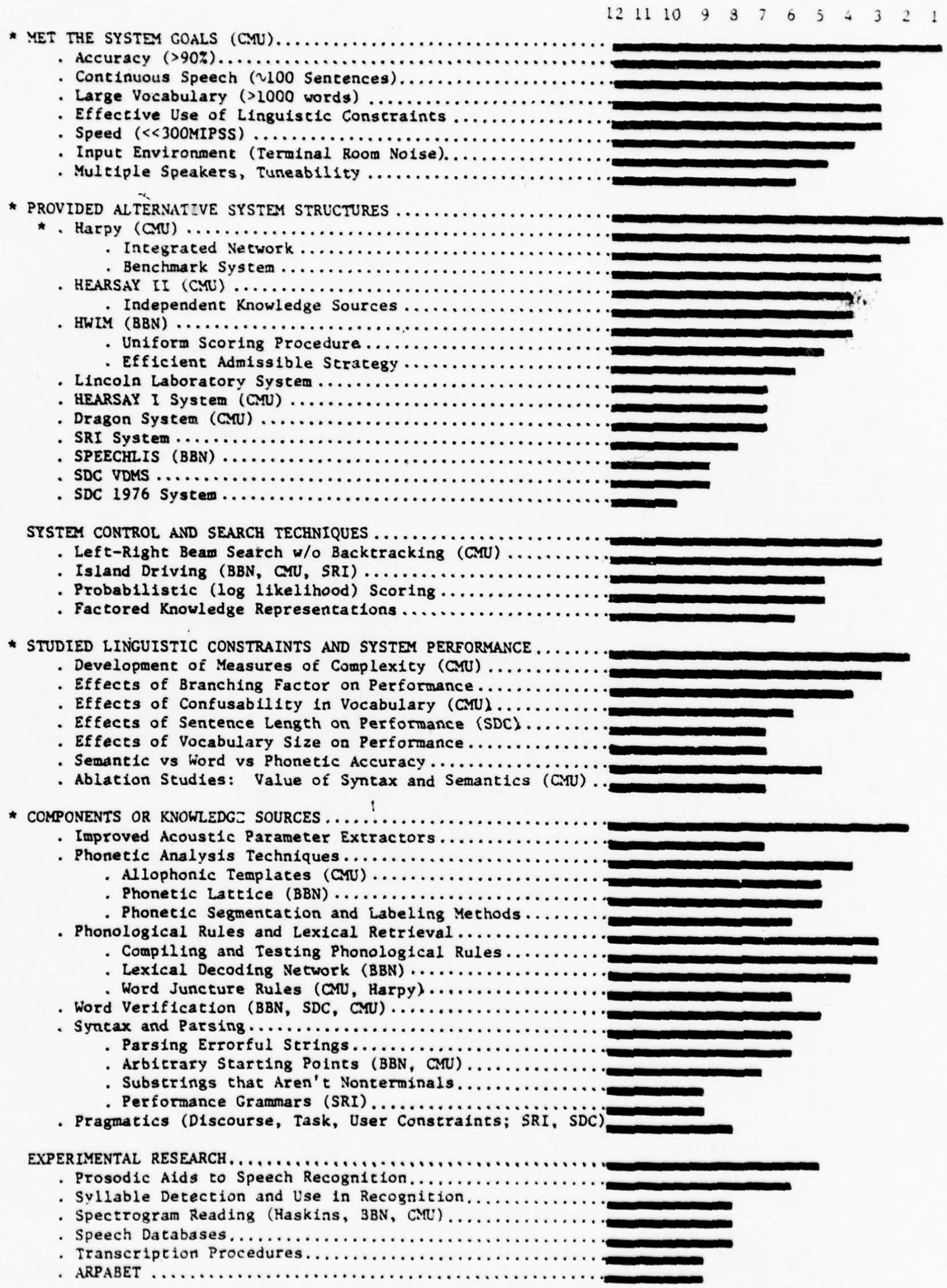


Figure 2-3. Primary ARPA SUR contributions and their relative significances. (1= most; 12= least significant of the notable contributions; *= highly significant contributions of score 3, 2, or 1)

1978 IBM CONTINUOUS SPEECH RECOGNIZER

Figure 1-3. Some highlights in the history of speech recognition.

6

training for each talker. It does not lend itself to easy incremental modification or improvement (such as adding new words or structures). Its extendability to more complex tasks is uncertain, and the "habitability" (or ease of learning) of its finite state language is in question.

Among all the other system structures investigated in the project, HEARSAY II and HWIM offer two notable contributions. HEARSAY II offers a promising contrast to Harpy, in its structure of independently operating knowledge sources which can be separately tested and evaluated, and modified. It is probably the easiest structure to use for testing out new ideas within old components, or adding new knowledge sources, such as prosodics or user models. HWIM, on the other hand, offers some excellent traditionally important knowledge sources that are relatively fixed in form (and somewhat more difficult to alter), but which may be used in a variety of control strategies, including "admissable" strategies that are guaranteed to find the best possible interpretation of the detected phonetic string. As noted in the latter part of Figure 2-3, among HWIM's notable contributions in components and recognition knowledge sources are its lexical decoding network, its phonetic lattice, its parametric word verifier, and its uniform (log likelihood or probabilistic) procedures for scoring hypotheses at all levels in the analysis,

It should be noted that HWIM, and to some degree, HEARSAY II also, were not carefully adjusted, tested with extensive data, or adequately analyzed to determine their weak or strong components. As HWIM developers have noted (Wolf and Woods, 1979), HWIM was just barely operational at the time of the final demonstrations, with half the vocabulary totally new to the system, some components totally untested and others not sufficiently adjusted, and the promising "shortfall density" strategy not operating during the main tests. While Harpy's success illustrates the value of freezing components and strategies early enough to permit extensive testing and "fine-tuning" of the system and its components, the final performance results for HEARSAY II and HWIM cannot be taken as accurate indicators of their ultimate potentials. Extensive performance evaluation studies were proposed, and still might be appropriate to undertake. If so, it would be useful to comparatively evaluate the HEARSAY II and HWIM systems and their individual components, perhaps by testing on a common task. Documentation of such complete performance evaluation could prove useful to developers of future speech understanding systems.

Earlier systems developed during the project became "throw-away systems" that guided the design of the final systems, but did not generally represent the best candidates for use in future speech understanding work. HEARSAY I (like HEARSAY II) still seems to be valuable for its ability to readily permit insertions of new components, or deletion of components to establish the specific contribution of each component. The Lincoln Laboratories intermediate system was impressively successful, and influenced HWIM substantially. Dragon was the predecessor of Harpy and also influenced the later designs of the IBM systems (Bahl, et al., 1978).

2.3.2.3 System Control and Search Techniques - Significant contributions were made in experimenting with various system control and search strategies. Harpy's strict left-to-right search without backtracking helped limit the combinatorial expansion of alternative word sequences to hypothesize, while HEARSAY II and HWIM's island driving from abritrary starting points in the utterance tended to create combinatorial explosions in alternative extensions of previously hypothesized word sequences. It appears that island driving, or middle-out analysis of an utterance, can be effective only if the islands are highly reliable, such as with multiple-word islands, or if the control strategy can effectively limit combinatorics.

Figure 1-3. The focus was on bringing advances in artificial intelligence and computational linguistics to bear on the task of having the machine comprehend the full linguistic structure, and producing the intended machine response appropriate to the meaning of a sentence or discourse (Klatt, 1977).

The systems converged on the idea of probabilistic (or log likelihood) scoring of hypotheses, and BBN developed a valuable uniform scoring procedure for relating assessments of hypotheses at all levels in a system's analysis. However, there was general agreement that a priori probabilities, whereby alternative hypotheses are selected on the basis of their frequency of occurrence in the language (or in previous data collections), should not be used, since they can cause the system to fatally reject correct but unexpected utterances or words. Unless one collects hours of statistical training data, such as the IBM system does (Bahl, et al., 1978), a priori probabilities can be in error due to fortuitous absence of acceptable utterances in the training data.

Another prominent trend from the project was the growing use of "factored knowledge representations" (cf. Wolf and Woods, 1979), whereby high-scoring knowledge from early hypotheses is used (at the same or other levels) in efficiently testing later hypotheses that also suggest similar local interpretation. (See Section B-1.2 of Appendix B.)

2.3.2.4 Studies of How Linguistic Constraints Influence Performance - A primary contribution of the ARPA SUR project was its simplifying of the recognition task by constraining it markedly via syntactic, semantic, and task constraints. This is comparable to lexical constraints in a small-vocabulary isolated-word recognizer. Because continuous speech recognition is a multiple-dimension problem, it can be constrained in many ways, and one question addressed by ARPA SUR work concerned which constraints were most effective in improving performance. System performances were found to be more closely associated with the branching factor of the language than with any other system variable. Harpy thus cannot be unequivocably appraised as "winner", since its task was (on one measure) almost an order of magnitude easier than that undertaken with HWIM. Interesting studies also showed that the confusions among similar words in the vocabulary reduce system performance more than the numerical size of the vocabulary (Goodman, 1976). Length of utterance is also more likely to influence recognition than vocabulary size (Bernstein, et al., 1976; also Barnett, et al., 1979). Focus of attention is now properly on language complexity (measured by branching factor, or perhaps by "entropy" or "perplexity"), word similarities, and utterance length, not vocabulary size.

Since final semantic accuracy (as high as 95%) was much higher than might be predicted from the low phonetic identification accuracy (40-50%) and multiplicative errors in word identification, it is evident that linguistic constraints substantially aided recognition. This was also graphically illustrated from ablation studies (comparisons of results with versus without syntactic and semantic components). These studies support the basic principles underlining the ARPA SUR effort, that acoustic information alone is not enough, that multiple sources of knowledge are needed, and that accurate "understanding" is different from accurate phonetic or word sequence "recognition" (Newell, 1975).

It will be important in the future to extend these valuable contributions from complexity metrics, ablation studies, and various system performance metrics, to provide quantitative ways of assessing performance of alternative systems even when they work on different tasks and with different knowledge sources and control strategies.

2.3.2.5 Components or Knowlege Sources - Major improvements were made in many previous recognition procedures, including acoustic parameter extractors, detailed procedures for detecting and identifying various vowels and consonants, and improved methods for word matching, but Figure 2-3 lists only the most noteworthy new contributions. Formant tracking from LPC spectra, im-

• RESTRICTED SPEECH UNDERSTANDING "Tell me about Nixon and Watergate."

• TASK-INDEPENDENT CONTINUOUS SPEECH

"What is the big event in Anaheim?" . . . "How many aircraft carriers does Russia have?"

Figure 1-4. Forms of speech recognized by machines.

8

proved autocorrelation methods for pitch tracking, and accurate procedures for detecting syllabic nuclei were among the improved acoustic analysis tools. In phonetic segmentation and labeling, Harpy's 98 allophonic templates for classifying short sub-phonemic segments are considered a significant contribution, as is HWIM's phonetic lattice. However, we still do not know which of these (or other) phonetic analysis techniques is best for future work. Experts we have surveyed overwhelmingly agree that better acoustic phonetic analyses (or "front end" components) are high priority aspects in developing future speech understanding systems.

BBN's lexical decoding network is considered a substantial contribution to efficient phonological analysis and word hypothesizing, and is particularly noteworthy for its handling of influences of words on sound structures of neighboring words. Harpy's juncture rules perform a similar function, but are not well documented or readily learned by the developer or user of the system. The cooperative compilation and testing of a large set of phonological rules was a major contribution. Word verification (especially HWIM's parametric word verifier) is another interesting contribution.

In syntax and semantics, the most noteworthy advances were the ability to parse errorful strings, work both directions from arbitrary starting points in the structure, and parse word sequences (substrings) that do not constitute single nonterminals in the grammar. Prosodic aids to parsing are promising but largely untested additions. While interesting experiments and a few implemented procedures regarding pragmatics were introduced, much more seems possible in the areas of using discourse constraints, task constraints, and user models. The ARPA SUR project served more as a user of higher-level linguistic components than as developer of new linguistic ideas and techniques.

2.3.2.6 Experimental Research - Experiments at BBN, Haskins Laboratories, and CMU showed that humans can transcribe speech from reading spectrograms, without knowledge of the linguistic context, and will be accurate about categorizing 70% of the phonemes in the utterances, and can be aided in identifying words in the spectrograms by computer-assisted survey of words of similar sound structure. Klatt and Stevens (1972) found that stressed cardinal vowels (/i, a, u/), prestressed consonants, and nasals that are not in consonantal clusters were most reliably identified. Sperry Univac also showed that vowels and obstruents in stressed syllables were most reliably categorized by several available machine transcriptions of speech. Vowels and syllabic nuclei were shown by such studies to be very reliably detected. Other experiments (principally at Sperry Univac) showed the importance of stressed syllables, intonational phrase boundaries, and rate of speech in recognition. Methods were developed for locating stressed syllables automatically, and extensive experiments showed that listeners could consistently decide which syllables were stressed. Valuable regularities were found in stress assignment for various word categories and phrase structures, and procedures were outlined for using such prosodic cues in speech recognition.

An agreed-upon ARPABET for transcribing and mechanically classifying speech sounds into phoneme-like units was developed, and procedures for orthographic, phonemic, and phonetic transcription were defined. Many large speech databases recorded and used during the project should be suitable for future research and system developments. Unfortunately, these databases, and the many computer algorithms for various aspects of recognition, have not been cataloged or made readily available to other researchers, and may be particularly difficult to obtain and use now that most (our survey shows 65%) of the ARPA SUR workers are dispersed and not working on speech understanding or allied projects. The only contractors that still have most of their key ARPA SUR personnel are BBN and SRI.

ANALYSIS

Figure 1-5. Models of "recognition" and "understanding"

9

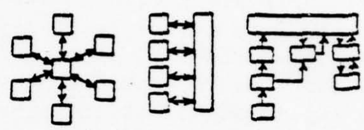
2.3.3 Advances in the State of the Art

Figure 2-4 summarizes some important aspects of the advances in the state of speech recognition technology which the ARPA SUR project produced (cf. Figure 2-1). Our discussion here closely parallels the description in Section 2.2.1 of the 1971 state of the art, so that the reader can readily make comparisons.

- ▣ COMMERCIAL ISOLATED WORD RECOGNIZERS
  - ▣ SMALL VOCABULARIES (10-200 Words)
  - ▣ ACCURACY = 99%
  - ▣ 6 COMMERCIAL SOURCES, REDUCED COSTS
  - ▣ HIGH-QUALITY SPEECH OR TELEPHONE
- • MORE DATA ON ACOUSTIC CHARACTERISTICS OF SENTENCES
- • COMPILATIONS, TESTS, AND USAGE OF PHONOLOGICAL RULES
- • USEFUL MEASURES OF COMPLEXITY AND PERFORMANCE
- • ADVANCED SCORING PROCEDURES AND ADMISSABLE STRATEGIES
- ▣ PROTOTYPE RECOGNIZERS OF WORD SEQUENCES AND SENTENCES (Bell Laboratories, IBM, NEC, Sperry Univac)

KEY: • ARPA SUR CONTRIBUTIONS
▣ Non-ARPA CONTRIBUTIONS

- • HARPY CONTINUOUS SPEECH RECOGNIZER 1,000 words, BF=33, Accuracy=95%
- • SEVERAL MULTIPLE-KNOWLEDGE-SOURCE SYSTEMS
  - • HEARSAY I, II; HWIM; others



  - • Systems With Acoustic Phonetics, Phonology, Word Matching, Syntax, Semantics, Pragmatics, and Some Prosodics
  - • Clear Evidence About Utility of Syntax and Semantics
  - • Speech Parsers Handle Errors and Arbitrary Strating Points
  - • Guidelines for Prosodic Aids to Recognition
  - • Large Lexicons That Handle Pronunciation, Variablities and Word Boundary Effects
  - • Improved Phonetic Unit Detectors; Allophonic Templates and Phonetic Lattice

Figure 2-4. The state of ASR after the ARPA SUR Project (1976).

A primary contribution of the ARPA SUR project was in showing the feasibility of continuous speech recognition, so that now restricted forms of sentence understanding have attained a high (95%) level of accuracy comparable to that in the prototype word recognizers of 1971. Vocabularies of 1000 words have been attacked, and, what is more, experiments showed that large vocabularies did not lead to the expected combinatoric problems of major magnitude. System performance is only slightly degraded in expanding vocabularies from 200 to 1000 words. Several systems now have been built that use multiple forms of incomplete knowledge and any of several interesting control structures, to handle spoken sentences. In addition to previously available and improved acoustic phonetic and word matching procedures, ARPA SUR systems incorporated phonological rules, some prosodic features, word verifiers, speech parsers that could handle errorful hypothesized word sequences and arbitrary starting points for analysis, and semantic and pragmatic analyzers.

Not only is there now at least one adequate system (Harpy) and a few other partially-successful and promising speech understanding systems (particularly HEARSAY II and HWIM); even more important is the advanced knowledge attained about alternative system designs and improved components or knowledge sources. Promising ideas and techniques were developed in each aspect of system operation, and some experimental evidence is available about the effectiveness of alternative analysis techniques. A few experiments also advanced our corporate knowledge about the acoustic characteristics of phonetic segments and prosodics in spoken sentences. Phonological rules have been gathered from the literature and from new studies, and implemented and tested for their effectiveness in aiding speech understanding. Several effective methods have been devised and tested for using phonetic and prosodic data and phonological rules along with higher-level linguistic constraints. Syntactic parsers now can deal effectively with errorful input strings, and "island driving" ("middle-out" parsing from arbitrary starting points) is possible, though experiments showed that such island-driving is likely to yield combinatorial explosions of alternative word sequences unless the islands are very reliably identified and are extended in well-controlled ways. In addition to traditional bottom-up and top-down control strategies, we now have several effective new strategies, including the Harpy integrated network with beam search, the HEARSAY II "blackboard" model of an "hypothesize and test" strategy, the various BBN strategies, and many others experimented with at BBN, CMU, SRI, Lincoln Laboratory, and SDC.

As researchers had expected, final accuracy of semantic understanding (e.g., 95% in Harpy) was clearly shown to be different from accuracy of phonetic segmentation (42%) and word matching (97%). Thus, phonetic classification accuracy could be as low as 42% and still allow adequate accuracy of semantic understanding. Studies with HEARSAY I showed that adding syntax to acoustic phonetic recognition procedures increased the semantic accuracy by 25% (from 40% to 65%), and the further addition of a semantic component added another 25% improvement. Evidence is now in that heavy syntactic and semantic constraints substantially aid recognition. The ARPA SUR work added several new measures of task complexity other than vocabulary size or number of words in a sentence; the branching factor was shown to be a good (but not totally adequate) measure of language complexity, and word confusability and entropy were also introduced to help evaluate systems. Finally, speeds of complex speech understanding systems are now known, and are expected to soon approach desirable real-time operation, even on moderate-sized computers.

While other systems have since been purported to meet or exceed the 1976 ARPA SUR capabilities, it is good to recall that the growth of other speech recognition projects after the beginning of ARPA SUR is not totally coincidental. That, too, is a fulfillment of an initial ARPA SUR goal, which was to stimulate other projects in continuous speech recognition (Newell, et al., 1971; p. 3). Recognition of continuous speech no longer is an unattained hope; it is an available reality in the form of HARPY-like systems, as well as in other restricted systems like those that recognize connected word sequences at Nippon Electric Company, IBM, Bell Laboratories, and Sperry Univac.

like acoustics, phonetics, lexical processing (word matching), prosodics, syntax, semantics, and pragmatics, plus appropriate control structures. Only the acoustic phonetics and word matching procedures were actually functional in available recognizers.

3. SUMMARY OF CURRENT (1978) TECHNOLOGY

Up to this point, we have considered why speech recognizers may be useful, briefly reviewed the 26 year history of speech recognition, and studied the impact of the large ARPA SUR project. Now where does all this leave us currently? We shall consider three aspects of current technology: (1) what can be obtained right now in practical isolated word recognition (Section 3.1); (2) how those recognizers are being used (Section 3.2); and (3) what development projects are being done to advance speech recognition technology (Section 3.3).

3.1 What Can You Obtain Right Now?

After 26 years of increasingly more intensive work on machine recognition of speech, there is now a well-established technology for inputting highly-restricted spoken commands. Currently one can buy any of several isolated word recognition devices, as shown in Figure 3.1. These current recognizers range over a broad spectrum from cheap hobbyist subsystems to expensive accurate systems ready for fully effective use in field applications. Heuristics, Incorporated makes a $299 hobbyist's "Speechlab," which provides only an acoustic front-end of a recognizer and requires a separate computer, although Heuristics offers software and ideas for small-vocabulary word recognition which are purported to provide over 95% correct word recognition (Enea and Reykjalin, 1978). Heuristics also offers manuals that introduce basic concepts of speech recognition and explain their recognition methods. Over 500 Heuristics Speechlabs are said to have been sold. Phonetics, Incorporated offers their SR-8 "stand-alone" recognition system for $550, which is capable of speaker dependent recognition of 16-word vocabularies, or may be used with a separate computer for larger vocabularies. This system functions as an acoustic pattern classifier, processing the total spoken word as a single complex (unsegmented) pattern, and benefits considerably from use of a time normalization algorithm. It is purported to provide 98% recognition with a 2% reject rate for experienced users (Hitchcock, 1978). A higher performance Phonics system is expected to be available in late 1978. Centigram's "Mike" system is almost an order of magnitude more expensive at $3,000 to $5,000, and is intended to be used with the ADAM computer. We know of no public claims of performance figures for this recognizer, but it has been publicly demonstrated (e.g., at the 1978 National Computer Conference). Perception Technology, Incorporated is also reported to have offered speech recognizers, but we have no information on them.

More expensive (but generally more accurate and versatile) total recognition systems are available from Interstate Electronics, Dialog Systems, Threshold Technology, and Nippon Electric Company. Interstate Electronics now manufactures a version of an earlier "VDETS" recognizer developed by a long-established speech recognition branch of Scope Electronics. Several VDETS (Voice Data Entry Terminal System) terminals and other Scope products have been used or tested by government agencies, and a Scope Electronics Voice Command System has been used and modified by NASA Ames Research Center (Coler, et al., 1978; cf, also Section 3.3). The Interstate Model 1832 speaker-dependent VDETS is intended for moderate size vocabularies (up to 300 words) and costs $18,750. Another expanded model for 900-word speaker-dependent recognition (or 4-user, 250-word recognition) costs $22,500. Interstate also

offers other optional devices and services to integrate voice data entry into a total applications system. For many practical applications, the effort and costs in developing a full application-oriented system far exceed the initial cost and trouble of obtaining a speech recognizer. Interstate, Dialog, and Threshold have offered total systems for voice input of data or commands, which represents a major step above initial recognizer capabilities alone.

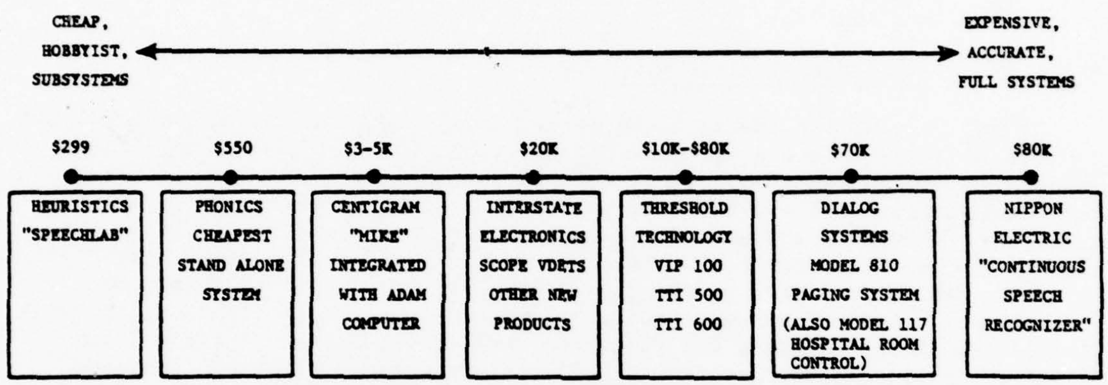


Figure 3-1. The Spectrum of available commercial speech recognizers.

Dialog Systems Incorporated has provided several recognizers for various applications, including the 99-word vocabulary Dialog 117 System for voice control of hospital room environmental conditions, for immobilized patients to control bed motors, lights, TV, nurse calls, etc. Dialog's standard Model 810 Voice Input Terminal is available for $70,000, for 12- or 31-word paging applications. This terminal, like many of Dialog's systems, works over telephone lines, and provides voice answer-back. Related Dialog Systems devices have been tested by government agencies with what was reported to us as considerable satisfaction.

Threshold Technology Incorporated has been the leading source of speech recognizers, selling well over 200 recognition systems since they offered their first project in 1972. The Threshold VIP 100 was tested and used in many commercial and governmental applications, and some are still in use or have been modified for advanced applications of isolated word recognition (Connolly, 1978). The VIP 100 was superceded by the Threshold 500, introduced in 1975, and the Threshold 600, introduced in 1977. The Threshold 600 is a CRT-compatible voice data entry terminal that looks to a computer just like a standard CRT terminal. Threshold Technology voice input systems cost from $10,000 up to $80,000 or more, depending greatly upon the customer's needs. Most of the successful applications of speech recognition in various commercial (and governmental) systems have been based on use of Threshold Technology devices, which provide speaker-adaptive, small vocabulary recognition

show the feasibility of understanding continuous speech, and each of the final systems was demonstrated (with varying degrees of accuracy) by handling

with accuracies exceeding 99.5% in the best of cases. Threshold Technology's systems generally work with single speakers or a few speakers, using head-mounted close-talking microphones, but some tests have been done with speaker-independent recognition (e.g., Scott, 1976) and telephone channels.

Recently Nippon Electric Company has announced a recognizer that reportedly can handle either 60 to 120 isolated words or else connected digits or word sequences, with over 99.5% correct recognition. The Nippon Electric DP Voice Recognition System uses a two-stage dynamic programming algorithm for time normalization, and uses distance measures between incoming spectral data and stored spectral templates. It is projected to cost around $80,000.

All of these available systems are primarily intended for speaker-adaptive use, requiring a moderate amount of training for each new talker. These commerical devices have done much to advance the credibility of voice input and to promote further concern with improved speech recognition capabilities. Current efforts are directed towards bringing the costs for such devices down to under $1,000, plus improving the ability to use telephone speech, and eliminating the need for each talker to train the system to the peculiarities of his or her voice. Industry, government, and researchers are now concerned with defining adequate standard tests to comparatively evaluate the available systems, particularly since the number of commercial sources of such devices has more than tripled in the last few years, and the systems being offered vary greatly in accuracy and price.

3.2 What Are Recognizers Currently Used For?

Around 300 isolated word recognizers (besides over 500 hobbyist devices) have been sold in the past six years, with roughly half of them now in daily operational use, and others being tested in experimental government facilities and research laboratories. There have been many satisfied customers, who report 50% to 90% reductions in manpower needs, time savings from 30% to 95%, increased accuracy of data entry, and higher user morale. Figure 3-2 illustrates some of the commercial and military applications. Some applications have involved package sortation systems (such as at UPS, S.S. Kresge, the U.S. Postal Service, and airline companies), where the operator's hands are busy controlling packages on a conveyor belt system and orienting the packages, while his spoken commands can simultaneously control on which belt each package should go. Other applications involve inspecting TV face-plates at Owens Illinois, compressors at Tecumseh Projects, pull-top can lids at Continental Can and Reynolds Metals, automobiles on assembly lines at General Motors and Chrysler Motors, etc. Machine tools like lathes and complex drilling machines can automatically be controlled by voice with highly sophisticated Voice Numerical Control (VNC) systems such as have been installed at Heat Controls, Inc., Calabrese and Sons, Joseph Moreng Iron Works, Purcell Manufacturing Co., Diversified Manufacturing Co., and other metal working facilities. Also, there is considerable interest in telephone banking and voice authorization of credit card transactions. One of the most successful recognizers ever developed was a system for securing personnel access to a computer facility, at Texas Instruments (Doddington, 1976). Other applications include voice actuated wheelchairs and hospital room environmental controls that are voice operated. The potential commercial applications of recognizers seem to be rapidly expanding.

the problem of handling multiple speakers was directly addressed, with con-
siderable success, at least for Harpy. In contrast, after 15 years of various projects, IBM and other long-term contributors to the field were still handling the speech of only one male speaker in 1978 (Bahl, et al., 1978). Speaker

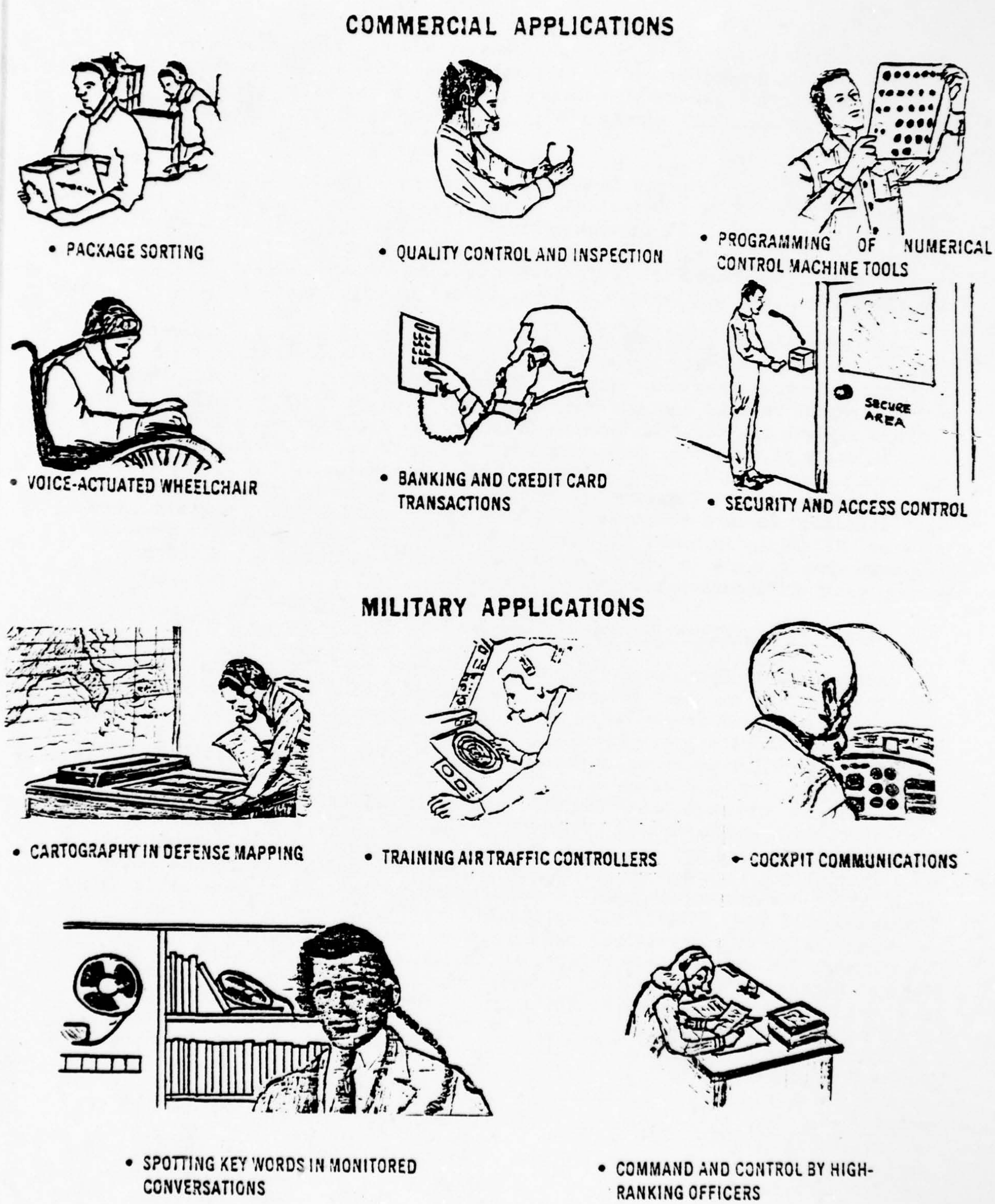


Figure 3-2. Commercial and military applications for speech recognition.

The military has been the primary source of development funds in speech recognition, and as shown in the bottom of Figure 3-2, the current DOD applications include cartography or map making (Beek, et al., 1977; Goodman, et al., 1977), computer-assisted training of skilled communicators like air traffic controllers (Breaux, 1978; Grady, et al., 1978), recognition aids in airplane and helicopter cockpit communications (Curran, 1978; Huff, et al., 1978), monitoring of large communications systems to automatically detect important conversations (Beek, et al., 1977), and natural forms of command and control operations by high-ranking officers who would like to speak natural commands to a machine without middlemen or complex input devices. More will be said in Section 6 about DOD applications for recognizers.

Other systems are being tested by the Federal Aviation Administration, the Veterans' Administration, and the NASA Ames Research Center, for air traffic control, voice-controlled wheelchairs hands-off control of hospital room environmental conditions, and simulations of pilot communications in helicopters and aircraft.

One marketing consultant (Michael Nye, 1979) predicts that in the next ten years about 2.5 million speech processing units will be sold, for a total market of about 4.8 billion dollars. At least one third of this market is expected to be in isolated word recognizers. Our survey of experts in speech recognition work (Appendix B) shows expected sales that can be projected to similar figures. With the expectation that accurate isolated word recognizers will come down in price within the next two years, from their current several tens of thousands, to well under $1,000 each, we can predict a rapidly expanding market, and a growing variety of applications.

3.3 What Development Projects Are Currently Active?

Currently, a number of developmental projects are being conducted at Bell Laboratories, IBM, ITT, Logicon, Nippon Electric Company, Sperry Univac, and Texas Instruments. CMU is still continuing work on extensions of Harpy, while other advanced research projects, especially including speech understanding work, are being conducted in Japan, France, Germany, Italy, and other countries.

We can only briefly outline these projects here, but the reader is referred to the publications and reports from these groups, and recent review articles (Reddy, 1976; Wolf, 1976; Lea, 1979c) for further details. Reviews of current work in Western Europe, Japan, and Poland are presented in several forthcoming articles (Haton, 1979; Wakita and Makino, 1979; Jassem, 1979).

3.3.1 Bell Laboratories - The latest in Bell Laboratories' long history of speech recognition studies have been concerned with (1) sophisticated isolated word recognizers using linguistic and task constraints on sequences of commands, and (2) several recognizers of digit strings and restricted word sequences spoken continuously (cf. Flanagan, et al., 1979). Itakura's (1975) recognizer was expanded upon and tested with an 84-item flight information vocabulary, yielding a median recognition accuracy of 91.6% that could be improved to 98.5% by a three-pass analysis to decide on the identity of more difficult words. Another system (Sambur and Rabiner, 1976) used specific sound structure features of each spoken digit to do 97% correct isolated

digit recognition for 10 speakers, and 94% for 55 speakers (30 female, 25 male) who had no prior experience with the system and who spoke in a noisy computer room, with a fairly low quality microphone.

A connected digit recognition system based on the isolated digit recognizer was developed, and tested in a speaker-dependent form. Average recognition accuracy was about 99% on strings of three digits, spoken by six speakers. For 10 new speakers who did not train the system, average accuracy was 95%.

Other recognizers developed at Bell Laboratories include a yes-no recognizer (which has as yet made no errors), and an interactive airline flight information system which has a finite-state syntax analyzer and a vocabulary of 127 words (Sondhi and Levinson, 1977). The flight information system was tested with seven speakers talking over dialed-up telephone lines, with a median rate of 88% for 10 speakers. Later, performance was improved into 96% correct recognition of name strings.

In addition to such excellent usage of grammatical constraints in recognition systems, Bell Laboratories also has developed a valuable "entropy" measure of grammatical constraint, for assessing the complexity of various recognition tasks (Sondhi and Levinson, 1977, 1978).

When one considers the excellent results and advanced ideas being developed at Bell Laboratries, one can conclude that practical accurate isolated word recognition has arrived, practical digit string and word sequence recognition is imminent, reasonable applications can be found for limited speech recognizers, and researchers are starting to get a handle on how to construct and evaluate an expanding array of recognition systems.

3.3.2 Carnegie-Mellon University - Harpy is being implemented for rapid processing on a PDP-11 mini-computer, allowing transportability to various groups interested in Harpy-like systems and extensions to Harpy. CMU has also cooperated with a graduate student in Texas, to help develop prosodic aids for a version of the Harpy recognizer. Lowerre and Reddy (1979) have offered suggestions for improving Harpy's ability to handle new words, new structures, and new tasks, as is discussed further in Section 2 and Appendix B. Apparently no further work has been done on a HEARSAY-type system.

3.3.3 The IBM System - The largest current effort in continuous speech recognition within the United States is being pursued at the IBM Thomas J. Watson Research Center. A very complete description of the system has been published by Jelinek (1976), but a brief statement will be given here. The speech recognizer is essentially a two-component system, involving an acoustic processor and a linguistic decoder. The input to the acoustic processor is the speech signal and the output is an estimated phonetic string. The linguistic decoder receives this phonetic string as an input and produces a printed output of words which hopefully represent the original utterance.

Originally the acoustic processor performed phonemic classification of individual spectra and then the speech was segmented into phonemes for the final output. More recently, the phone segmentation and labelling was eliminated; instead, a sequence of centisecond labels from a 33-phone alphabet is outputted to the linguistic decoder. The advantages given for the

centisecond-level modes are that information related to phone length is made available in a form usable by the models in the linguistic decoder, that more of the important information is preserved, and that segmentation and labelling decisions are delayed until decisions can be made by the linguistic decoder.

There are several subcomponents of the linguistic decoder. These include a phonemic dictionary, statistical phonological rules characterizing the speaker, and a statistical performance characterization of the front end. The statistics are obtained through an automatic iterative training technique. They are used in assigning likelihoods to hypothesized sentence fragments during decoding. Two types of decoding algorithms have been used, a "stack decoder" and a "Viterbi decoder", with the latter currently being applied to the centisecond acoustic states. The final output of the decoder consists of the most likely word string.

When this IBM system was trained to the voice of a particular speaker, it recognized seven-digit telephone numbers correctly 96% of the time, with a better than 99% per-digit accuracy. For a small finite-state language called the "New Raleigh Language", the phone-level model gave 73% sentence recognition, with 96.4% of all the words properly identified. The centisecond-level model showed improvement over the phone-level model on the same database, giving 95% sentence recognition, with 99.4% correct word identification (Bahl, et al., 1978). The system was also tested on continuous speech (from one speaker) of approximately seven-word length-sentences (namely, the task used previously to test Harpy), with a vocabulary of 1,011 words and with severe grammatical constraints. For this Harpy task, the centisecond-level model recognized 99% of all sentences correctly, with 99.9% correct word identification. On an entirely different corpus (laser patent texts) which had none of the grammatical constraints of the above-mentioned tasks, but allowed fully natural discourse with a 1,000-word vocabulary, results were obviously not as good. On a test set consisting of 20 sentences having a total of 486 words, there was only 66.9% correct word recognition (Bahl, et al., 1978).

While IBM work has been characterized by extensive use of statistical analyses and operation with a single speaker, current work is directed at handling more than one speaker.

3.3.4 ITT - A new project in speech recognition was instituted at ITT's Defense Communications Division in 1977, and while no complete recognition system has yet been developed, the researchers led by White and Sambur (1979) are exploring speaker independent speech recognition, telephone bandwidth and noisy speech, and low cost speech processing hardware. Isolated word recognition, word spotting, and talker identification are among the intended application areas, and prominent among the methods being investigated are dynamic programming and optimum distance (or "similarity") measures.

3.3.5 Logicon - Several studies have been conducted at Logicon, Incorporated, concerning the feasibility and utility of speech recognition for such tasks as the Automated Adaptive Flight Training System, and training air traffic controllers for the ground controlled approach system. Another task attempted was recognition of voice commands of a conning officer aboard a ship, for which simulations showed 90% correct recognition of a small vocabulary of 63 navigation commands (isolated phrases) spoken by several talkers within a task-dictated syntax. From 1973 to 1977, Logicon used commercial (Threshold Technology, VIP 100) isolated word recognizers coupled to their computer

facility, and in 1977 they began developing a system (called "LISTEN") for real-time recognizing of connected speech with small vocabularies. This LISTEN system uses a Markov model like Harpy and the IBM systems use, but they replace Harpy-like linguistic processing with faster mathematical and statistical analyses. No performance results are yet available for this developing system. Its capabilities are clearly limited to small vocabulary, finite-state languages such as are appropriate in restricted tasks like navigation commands and training of air traffic controllers.

3.3.6 Nippon Electric Company - Already mentioned in Section 3.1 and Figure 3-1 was the Nippon Electric "continuous speech recognizer", which extends isolated word recognition capabilities into limited forms of connected speech, such as digit strings. An unpublished announcement of their system reported 100% correct recognition of isolated words, and 99.9% for sequences of three digits, all spoken by five men. This is one of the most advanced dynamic programming analyzers available. Further development work is continuing at NEC.

3.3.7 Sperry Univac - A linguistically-based continuous speech recognition system is under continuing development at Sperry Univac's Defense Systems Division (Medress, et al., 1977; Medress, 1979). This system uses complex phonetic and prosodic analysis schemes and a lexical matching scheme based on matching analyzed phonetic sequences to lexically predicted sequences, and incorporates syntactic constraints to restrict the possible hypothesized word sequences. Expected phonetic confusions (represented in a "scoring matrix" and generative phonological rules) are used in guiding word matching procedures. Procedures are incorporated for assigning higher significance to "robust" or "highly reliable" phonetic units, and for allowing missing or extraneous segments and erroneous phonetic classifications to be overcome. The system reportedly achieved accurate phonetic analysis for vowels (94% correctly detected), sibilant consonants (92%), and retroflexive (r"-like") consonants, (85%), but had less success with palatal glides (30% correct) and voiced fricatives (24%). Overall system performance, in terms of percentages of connected word strings correctly recognized, was 94%, although those results included a large majority of utterances originally used for developing the system. The results for new test utterances alone was an average of 83%.

A slight modification of this system, with the addition of a word verifier, for getting a second chance to confirm the presence of hypothesized words, has been used for a word spotting system (Medress, et al., 1978). A "word spotter" must detect words in the context of any other words, without being sensitive to talker differences and channel distortions. Consequently, most early work on word spotting was concerned with detecting linguistically-invariant units such as phonemes or phonetic classes (cf. Lea, 1979c) and matching phonetic strings of the lexicon with those analyzed in the incoming speech. It was also acknowledged that detection of stressed syllables and other prosodic structures should also help spot the prominent content words in smooth flowing speech (Lea, 1973c; Medress, et al., 1978). An earlier study at Dialog Systems (Moshier, et al., 1977) had achieved over 90% detection of occurrences of the single word "Kissinger", with about 6 false alarms per hour, for tests with nine talkers, but performance dropped to about 70% detection for ten new voices. Doddington, et al. (1976), showed that shorter words (with less phonetic structure to check for) gave more word spotting errors than longer words, and Sperry Univac's study (Medress, et al., 1978) confirmed the value of long words with many robust sounds in them to aid in word spotting success.

3.3.8 Texas Instruments - A highly successful system for recognizing strings of six spoken digits, followed by a speaker verification procedure, has been developed at Texas Instruments (Doddington, 1976). The ultimate goal of this system is to verify from the speaker's voice that he or she is qualified to have access to a secure computer room. However, the more advanced version of the system involves the system first recognizing a sequence of digits, to help identify the individual's purported identification number. After the number is identified, then the speaker's specific voice characteristics are verified for granting access to the facility. The digit string recognition aspects of the system were shown to be almost 99% correct (using parity check digits to correct some recognition errors) when tested with 14 male subjects. Another "word spotting" capability was demonstrated in that about 95% of the occurrences of certain "passwords" were detected in free connected speech of eight talkers. An independent study by Mitre Corporation (Fejfar, 1977) concluded that the voice identification system was superior to other entry control systems such as handwriting or fingerprinting analysis, so that such voice entry systems should be a promising application of speech recognition technology in the future.

3.3.9 Recognition Work in Other Countries - Several other speech recognition systems (for both isolated speech and continuous speech recognition or understanding) are being developed around the world. For example, limited continuous speech understanding work is actively going on in France, Germany, Italy, and Japan. In Orsay, France, at Laboratorie d'Informatique pour la Mechanique et les Sciences de l'Ingenieur, Mariani and Lienard and their colleagues are working on a speech understanding system (called "ESOPE") that uses phonological, lexical, syntactic, prosodic, and semantic constraints to correct errors in acoustic phonetic analysis, in a manner similar to the ARPA SUR systems. They are currently investigating performances with several very limited tasks such as phrases composed of subject-verb complement constructions (21-word vocabulary with a branching factor of 5), digit strings of unspecified length, numbers (like "thirty four"), arithmetic expressions (42-word vocabulary, branching factor of 12), and phrases used in standard telephony (for which they are testing how performance varies with size of the vocabulary and branching factor). At Centre de Recherche en Informatique, Universite de Nancy, France, Haton and his colleagues are working on speech understanding in the MYRTILLE project (cf. Haton, 1979). In 1974-1976, a first version called MYRTILLE 1 was implemented using a top-down (syntax-driven) strategy. MYRTILLE 2 involves a new parser, prosodic segmentation schemes, new phonemic recognition procedures, and a versatile interactive system. A review of Western European work in speech recognition has recently been prepared by Haton (1979). It appears that the European scholars are continuing where the ARPA SUR project left off, but proceding with some caution, initially dealing with fairly restricted problems. One of the interesting highlights of European work is the use of syntactic pattern recognition schemes, which model speech waves by structural features or units whose composition and combinations are determined by syntactic rules (e.g., cf. Baudry and Dupeyrat, 1978; DeMori, et al., 1976). This method has been well known in other (visual) forms of pattern recognition work in the USA (Fu, 1974), but has had little attention in American work in speech recognition (cf. Lea, 1979, a,c).

In Japan, several recognition efforts are active (cf. Wakita and Makino, 1979, for a summary), but the primary work in speech understanding is at Kyoto University (however, work is also being done at Yamanachi University: Sekiguchi and Shigenaga, 1978). A multiple-knowledge-source system for

recognition of a small vocabulary spoken version of the "BASIC" programming language was developed by Niimi and his colleagues (1975), with some success. It achieved 47% correct sentence recognition in 1975, and has since been improved considerably. A more ambitious project was the LITHAN (LIsten-THink-ANswer) speech understanding system (Nakagawa, 1976), which used a 20-channel filter bank in an Acoustic Processor, plus a Phoneme Recognizer, Word Identifier, Word Predictor, Parsing Director, and Responder, to achieve 64% correct recognition of sentences of 4 to 24 words, with a 100-word vocabulary. A review of recent speech recognition work in Japan has been prepared by Wakita and Makino (1979).

Given this broad overview of current speech recognition technology, our next task is to consider the various areas or issues that currently seem to need further attention. The next section will summarize our conclusions concerning the "gaps" in current speech recognition technology.

4. GAPS IN SPEECH UNDERSTANDING TECHNOLOGY

Up to this point, we have shown that voice input to computers offers many advantages, has a spotty but advancing history, was significantly advanced by the ARPA SUR effort, and is currently a practical technology in the limited form of isolated word recognizers and a promising and active research and development field for more advanced continuous speech recognition. Here we consider the "gaps" in current speech recognition (or "understanding") technology, that will warrant further work. We consider various expert opinions about previous and current technology (Section 4.1), a comprehensive framework for defining gaps (Section 4.2), adequacies in current technology (Section 4.3), our assessment of gaps remaining after the ARPA SUR project (Section 4.4), and the general gaps between needed systems and current technology (Section 4.5), the most significant gaps in various aspects or recognition (Section 4.6), and opinions about future prospects for speech recognition (Section 4.7)

4.1 Expert Opinions about Previous Work and Current Technology

It seemed essential to consider the viewpoints of many experts as we assessed the total technology in speech recognition and understanding, and defined all "gaps" or high-priority problem areas in this complex field. Consequently, as noted previously, we visited the sites of participants in the ARPA SUR project, plus groups at Bell Laboratories, Dialong Systems, FAA/NAFEC, IBM, ITT, Interstate Electronics, NADC, RADC, Texas Instruments, and Threshold Technology. In our 30-page questionnaire sent to over 160 such workers, we solicited opinions about the ARPA SUR project, the current status of speech recognition technology, and future needs and trends. The detailed results of this survey are given in Appendix B. Sixteen of the 34 respondents to our questionnaire were participants in the ARPA SUR project, and eighteen were not. As shown in Figure B-4 of Appendix B, participants and non-participants agreed that the ARPA SUR project was important, needed, and ambitious, and that it produced a significant advancement bordering on a breakthrough. Participants considered the project well conceived and successful and that the original system design specifications were met, while non-participants generally did not. They agreed that while the meeting of original specifications was an important result, the advancement in specific aspects of recognition was more important. The independent research of the support contractors was considered more important by non-participants than by ARPA SUR participants, but both groups agreed that the idea of independent research efforts was a good one. Figures B-6 and B-7 show that the primary areas of contribution from the ARPA SUR project were agreed to be in system control strategies, segmentation and labeling, phonological rules, word matching and verification, syntax and parsing, and prosodics. Harpy and HEARSAY II were considered to be good systems but with limitations, while HWIM was considered to be a good system design despite its poor results. All the systems except HWIM were considered to have language-handling capabilities that were, on plurality vote, "far from habitable".

In assessing current technology, the expert respondents considered Harpy, HEARSAY II, the IBM system, and HWIM to be among the most relevant systems for future work. Non-ARPA experts responded more favorably than ARPA participants to the inclusion of IBM-like and "prosodically guided"

systems in future studies. The beam ("best few first") search strategy was considered best, and "island driving" was considered to be useful only if the islands were truly reliable. Primary among the knowledge sources or system components that warranted further work were the "front end" aspects of: segmentation and labeling, handling coarticulation and word boundary effects, use of phonological rules, word matching procedures, and prosodics. The need for better phonetic segmentation and labeling, or general acoustic phonetic analysis, was noted by every speech recognition group we visited around the country. When the respondents were asked another slightly different question concerning how strongly they agree with each of the choices of "the most significant 'gaps' (problem areas in need of further study) in speech understanding technology" acoustic phonetic analysis again tops their list of "gaps", along with prosodic cues to linguistic structures. Performance evaluation was next in importance, followed later in the list by a related topic of tuning the system with extensive data. The respondents also assigned fairly high significance to using linguistics to constrain ambiguties in the sequence of words. Near real-time processing and general scoring procedures were also included among the "gaps".

Other responses clearly showed that the majority of the experts believe that among the adequate aspects of recognition are: the acoustic parameters like formants and LPC analysis results, syntactic analysis that sharply constrains word sequences, semantic models that work together with syntax, and phonetically or parametrically based word matching procedures. Only the non-ARPA respondents considered syntax and parsing, semantics, and acoustic parameters to be of some priority.

Concerning the types of systems that are most needed now, ARPA SUR participants favored speech understanding systems and recognizers of connected word sequences, while non-ARPA participants favored more modest systems ranging from very low cost isolated word recognizers to recognizers of modestly-constrained connected word sequences. In general, the most needed systems (in descending order of average rank) were considered to be: linguistically constrained sequences of isolated words; digit string recognizers; recognizers of connected (formatted) word sequences; speech understanding systems that are much more powerful than Harpy; and low-cost isolated word recognizers.

Many other opinions about various technical details of current technology are presented in Appendix B, including recommendations about the best methods and most troublesome aspects of each component in a recognizer. In the remainder of Section 4, we shall present our own assessment of the adequacies and inadequacies in current technology, and the relative significances of the various "gaps" that must be bridged by further work. This presents a more cohesive picture than is possible from the consensus or conflicting views of various experts.

4.2 A Framework for Defining Gaps

It is questionable whether one can define gaps in a technology without defining the complete needs and current adequacies. Without committing ourselves to a specific speech understanding task or system structure (such as the classical "bottom up" hierarchical speech recognizer, the "top down" method, the "analysis-by-synthesis" scheme, the HEARSAY I system of independent knowledge sources, or the HARPY integrated network, or any other model), we suggest that total programs in developing speech understanding/recognition

systems involve the many aspects listed in Table 4-1. We have included not only technical dimensions of systems design and specific components, but also aspects of testing, evaluation, and project design that are important to defining good projects that will help fill gaps in the field. Supporting research and practical applications are also included as essential aspects of the process of developing a total technology in speech understanding.

The study report (Newell, et al., 1971) that originally defined the ARPA SUR project addressed itself to many of the design criteria listed under "Input Characteristics", "Task Constraints", and "Aspects of System Operation" listed in Table 4-1, and considered only the final semantic accuracy as a performance evaluation criterion. Throughout the project and its subsequent reviews, other issues have arisen regarding such aspects as databases, amount of speech processed, component evaluations, system comparisons, supporting research, and the best methods of project design and management. Other projects in this field have, of course, faced many of these issues previously. We consider next what aspects seem to be fairly adequately developed to date.

4.3 Adequacies in Current Technology

Only limited work has been done on the input characteristics shown in Table 4-1, and the only conditions under which fully adequate recognition is attainable today are high-quality speech from a speaker for whom the system is trained. For isolated word recognizers, telephone speech, noise, and many speakers of both sexes and multiple dialects do appear to be within reach, with a few successes already shown in previous studies. Continuous speech recognizers are clearly limited in their abilities to handle the more challenging input conditions.

Isolated word recognizers are adequate for small (30 word) vocabularies, and several studies show nearly adequate performance in using syntactically constrained sequences of isolated words. Digit string recognizers appear close to becoming effective commercial products, and not far behind are recognizers of strictly formatted word sequences. The feasibility of limited speech understanding has been demonstrated, and the successful HARPY system and its integrated network structure are being adequately applied, extended, re-implemented on a smaller computer, and used in developing a system for a practical task of computer-assisted training (Breaux, et al., 1978).

Significant advancement has been made in system organizations, so that now there are several promising system structures that could be used in future work. Syntactic models have proved effective, so that systems performed best with highly constrained syntax that truly restricted word sequences, rather than just providing a good description of the language. The other areas that appear to be sufficiently adequate to allow some future system development include semantics, some lexicon and word matching procedures, acoustic parameterization, and syllable nuclei detection. In the remaining (majority) of the aspects listed in Table 4-1, substantial further work is needed.

4.4 Specific Gaps Remaining After the ARPA SUR Project

Immediately after successfully providing an initial technology base for restricted forms of continuous speech recognition, the ARPA SUR project

TABLE 4-1. ASPECTS OF SPEECH UNDERSTANDING PROJECTS

INPUT CHARACTERISTICS	Transducer: Microphone, telephone, radio channel Channel Constraints: Bandwidth, noise, distortions Speaker Population: Number of speakers, sex, dialect, idiosyncracies
TASK CONSTRAINTS	Continuous speech or isolated words Language design and task complexity Training procedures, adaptation Requirements of real applications
ASPECTS OF SYSTEM OPERATION	Acoustic parameterization Segmentation and labelling Phonological rules Word matching and word verification Scoring procedures Prosodic aids to recognition Syntax Semantics Discourse analysis Task domain User models Statistics System structure and control Response mechanisms Generalized input-output systems
SYSTEM TESTING AND EVALUATION	Selection of databases Amount of speech processed Performances of system components Final system performance Comparisons of alternative systems
PROJECT DESIGN AND MANAGEMENT	Supporting research efforts Interim milestones Allowing time for system evaluation Back-up plans

ended in 1976. Extensive further work was still needed before practical commercial and military uses of speech understanding systems could become widespread. The limited ARPA SUR final system tests did not provide a total scientific evaluation, and the HWIM system in particular needed extensive further testing and adjustments. Many "gaps" remained in speech understanding technology. As we observed in Section 4.1, experts considered the following among the top-priority aspects of recognition that need attention (listed in descending order of priority):

- Acoustic phonetic analysis;
- Prosodic cues to linguistic structures;
- Performance evaluation;
- Using linguistics to constrain ambiguities;
- System tuning on extensive data;
- Fast or near-real-time processing;
- Scoring prodecures; and
- Phonological rules.

Figure 4-1 provides a more detailed list of prominent gaps in speech recognition, with our estimate of relative significances of gaps, in parallel fashion to the list of the ARPA SUR contributions in Figure 2-3. These areas in need of further work do not imply shortcomings or inadequacies in the ARPA SUR project, but indicate a composite grading of: (a) the importance of specific problems to overall system utility; (b) the degree to which needed performance must still be advanced, despite any previous advances coming from ARPA SUR work or other previous work; and (c) logical needs to answer some questions before others are pursued.

We can only briefly discuss these problem areas and their priorities here. In advancing system goals, studies with practical usage of isolated word recognizers indicate that 97% accuracy is usually needed, and at least that accuracy would probably be needed for the longer spoken sentences. Extensive testing and performance evaluating of total systems and their components is clearly needed for <u>all</u> systems, whether they be isolated word recognizers or speech understanding systems. Work with better use of linguistic constraints and habitable languages seems important, while speed, input channels, and multiple speakers don't appear to be limiting the utility of recognizers. There is no evidence that significantly larger vocabularies are needed.

A four-pronged effort seems to be called for, in which: (1) Harpy-like systems are developed and applied; (2) research systems with independent knowledge sources are developed and used as test beds for various new ideas; (3) experimental research is conducted on topics relevant to speech recognition; and (4) practical applications are addressed. Harpy can be applied to various restricted tasks such as voice entry of cartographic data (Goodman, <u>et al</u>., 1977), training of air traffic controllers (Breaux, 1978; Grady, <u>et al</u>., 1978), and various digit string recognition tasks. Extensions to Harpy that should really pay off include incremental compilation, automatic knowledge acquisition, and introduction of additional knowledge sources (such as prosodics). <u>Research systems</u> which permit the introduction and testing of new or improved components are also vitally needed, and should be permitted to test system features like uniform scoring, admissible strategies, and reliable island driving. Within those research systems, work

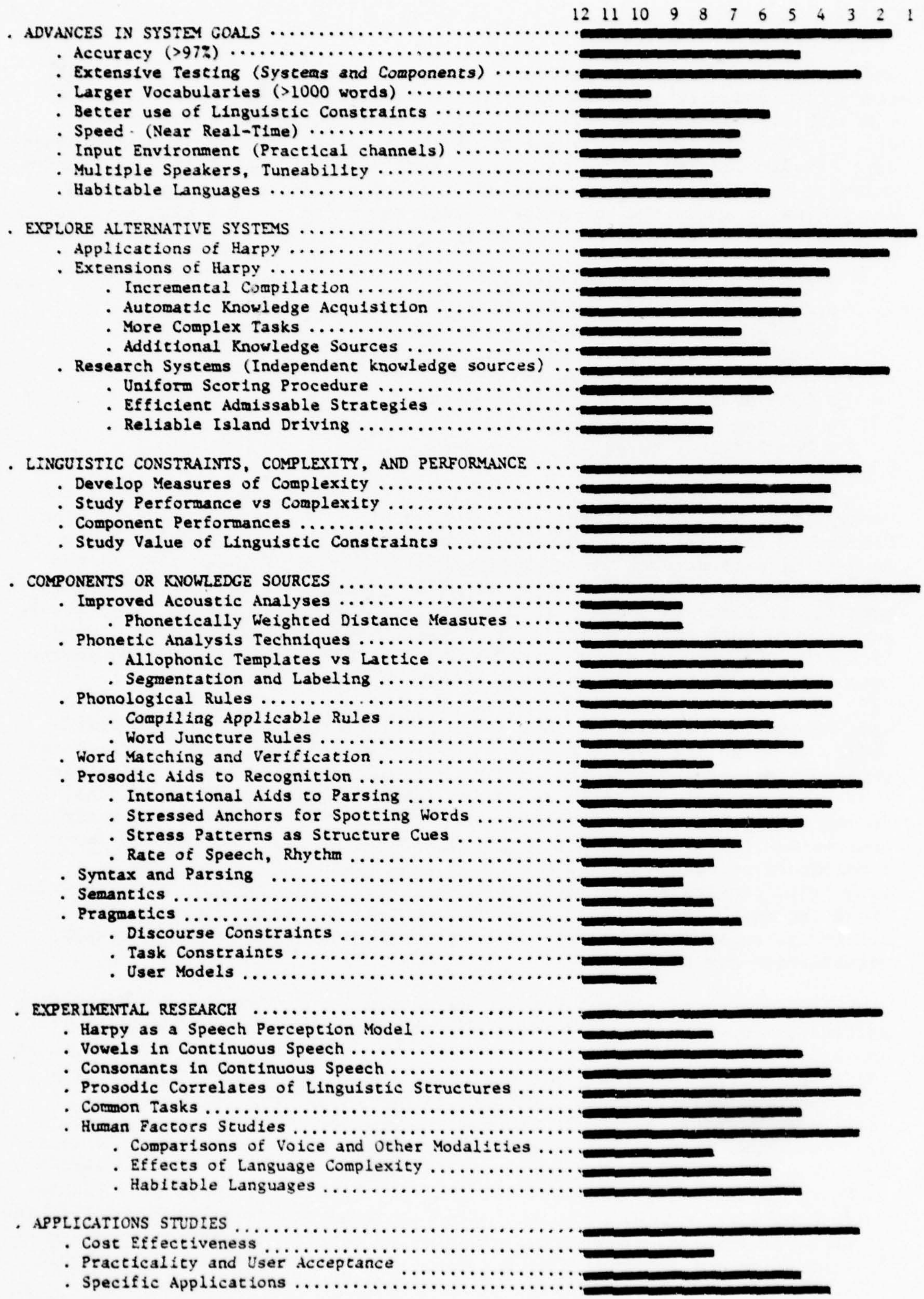


Figure 4-1. Primary areas in speech recognition technology needing further work.

is needed on phonetic, phonological, prosodic, and other types of components, with the specific tasks and priorities roughly as indicated in Figure 4-1. Coupled with recognition studies should be necessary experimental research on prosodic structures, human factors issues (especially concerning what makes a language habitable), comparative evaluation of systems with common tasks, and general characteristics of spoken sentences.

Finally, practical applications should be examined carefully. The initial ARPA SUR goals did not call for explicit applications to military (or commercial) needs, but, near the end of the project, pressures were on the contractors to show the near-term military relevance of their work. It is now appropriate to address operational needs for restricted continuous speech recognition (See Section 6).

System performance evaluation is an important adjunct to exploring alternative system structures. Total measures of recognition task complexity are needed, which more completely measure language complexity, sentence complexity, vocabulary confusability, confusability of phrases and their possible extensions, and other dimensions of difficulty such as speaker population, environmental conditions, etc. It currently is difficult to comparatively evaluate two or more alternative systems if they are not applied to tasks of equivalent complexity and similar or identical speech data, and we do not know how to decide whether system A which yields 90% recognition on a simple task is better or worse than system B which yields 50% recognition on a difficult task.

It is also important to determine the causes of recognition errors, and the weak and strong links in system operation. Systems need to adjusted, tested, and evaluated with extensive amounts of data. This is one reason why Harpy succeeded, since it was carefully adjusted with extensive speech data. Projects should be designed to allow time for such extensive tests, including over-all system accuracy scores and evaluations of the effectiveness and weakness of various system components.

The end of the ARPA SUR project also left undone several tasks necessary to make available to others the technology and scientific knowledge gained during the project. It would be valuable to compile lists, descriptions, and comparative evaluations of available speech databases, research results and facilities, reports, and system components or algorithms that resulted from the project. This could also be useful for the entire speech recognition field.

4.5 General Gaps Between Needed Systems and Current Capabilities

The specific gaps listed in the previous section primarily indicate the emphases that should be considered in future research and development projects. However, extensive thought must also be given to the general problem of how to best bridge the vast gap between isolated-word recognition and long range work on ambitious speech understanding (or sentence-recognizing) systems. Most practical applications of speech recognition have been demanding "off the shelf" completed systems, and thus have been trying to adapt their requirements to isolated-word recognizers. No absolute needs for continuous speech recognizers have been defined (Beek, et al., 1977), but this may in part be due to the available technology dictating to the application rather than allowing the applications to determine needed technology. The

desirability of continuous speech (due to increased speed and naturalness, reduced training demands on the users, etc.) may be evident, but the necessity or comparative cost effectiveness of versatile continuous speech recognition is not yet established. Part of the problem is to determine how much recognition capability is really needed. How versatile does the language of possible utterances have to be, and can one define a hierarchy of tasks of increasing complexity that can and should be handled by speech recognizers? Can systems be defined that are useful but have capabilites somewhere between isolated words and speech understanding systems? If so, what are those intermediate systems? Also, how acceptable is it to have systems that require extensive training, such as Harpy and some other systems need, and can systems be developed that require little training?

No systematic studies have been done to determine what accuracy (and other system features) it takes for user acceptance of a recognizer. For some situations, errors are so intolerable that feedback of recognition decisions is essential to correct occasional errors. Work still remains to be done (even with simple isolated word recognizers) on reducing the performance degradation that results from various levels of background noise, the effects of mircrophone characteristics and their physical mountings, and the difficulties introduced by physical stress (labor, heavy breathing, etc.) and mental stress (danger, critical operations, or being overloaded with things out of control).

With the advent of many commercial sources of word recognizers has come a growing demand for objective procedures for evaluating systems. Several manufacturers have mentioned to us the need for industry standards, such as general speech databases of 100 or more talkers speaking many instances of the digits and other vocabularies such as the alphabet ("A, B, C, ..."), the phonetic alphabet ("alpha, bravo, charlie, ..."), the names of the 50 states, or vocabularies relevant to important applications. The Speech Recognition subcommittee of the IEEE Machine Intelligence and Pattern Analysis Technical Committee of the IEEE Computer Society is currently endeavoring to define several general purpose databases that can be used to comparatively evaluate recognizers. However, just having individual manufacturers run the same speech data through alternative recognizers does not provide totally adequate "bench-marks" for assessing recognizers. The total task of interactions between unsophisticated users and voice entry systems needs to be tested in accurate simulations of field usage, which could perhaps best be carried out by an independent evaluation facility.

Workers in speech recognition would thus do well to consider the following general questions as they attempt to develop specific systems or generally advance the technology of speech recognition:

- Is there a real need for a speech recognizer in the intended application?
    - What advantages of voice input recommend its use in the application, in preference to other communication modalities?
- Can available "off the shelf" speech recognizers handle the conditions of the intended application?

- What form of recognizer is best used in this application?
    - Isolated words, recognized independently?
    - Isolated words in syntactically constrained sequences?
    - Digit string? (Fixed lengths or variable lengths of strings?)
    - Formatted sequences of words, spoken without pauses?
    - Sentences from a highly-restricted Harpy-like finite state language?
    - Sentences from a less-restricted speech understanding task?
    - Sentences unrestricted by task constraints?

Neuburg (1979) has suggested that at least some segment of the speech research community concerned with speech recognition should be addressing modest "next steps" in recognizer developments. This is in line with the concept of "bootstrapping" oneself up to better and better recognition capabilities, by careful stages of progression. There is an ever-present danger in such a cautious strategy that a dead-end may be reached, analogous to when one tries to get to the moon by climbing a tree. Ambitious projects which reach far, with the best of the available tools, such as the ARPA SUR project did with higher-level linguistic constraints, are also needed to complement cautious efforts on modest advances.

Neuburg also raised the provocative question (cf. Lea, 1970a,b) of whether work on speech recognition constitutes a "problem" (i.e., a real need for computer input capabilities) in search of an adequate "solution" (an accurate recognizer that meets the needs of the task), or a "solution" (someone's pet device) in search of a "problem" (a buyer). For isolated word recognition, the answer seems clear. Current workable systems constitute <u>solutions</u>, and they appear to be solving an increasing number of practical <u>problems</u>. For limited continuous speech recognizers dealing with digit strings or formatted sequences of words, several applications ("problems") seem to be calling for the imminent "solutions" represented by the successful prototype recognizers recently developed. For speech understanding systems of highly restricted form, the feasibility demonstrations of Harpy and other systems suggest the forthcoming appearance of practical "solutions" that will be seeking practical applications. Versatile speech understanding systems are not yet existent "solutions", and there are no "problems" unquestionably requiring their usage. While the burden of proof regarding the ultimate value of such versatile speech recognizers rests firmly on the proponents of such systems, the researchers and system developers cannot be expected to anticipate all the potential users to which an advanced system can be applied. There is some truth to the concept that <u>after</u> you "build a better mousetrap", <u>then</u> customers may subsequently come clamoring for it. Most users will not always see in advance why they "need" a better solution if what they already have (such as other input modalities or isolated word recognizers) are at least working somewhat adequately. Industry, for example, will rarely fund the research and development of new devices such as versatile speech recognizers.

5. RECOMMENDATIONS FOR ADVANCING SPEECH RECOGNITION

We finally come to the culmination of our work, in our conclusions about how to bridge the gaps in current technology and further advance speech recognition. We consider the consensus of expert opinions about another ARPA SUR-like project (Section 5.1), the need for coordination and ARPANET-like interactions (Section 5.2), some specific programs that should be undertaken (Section 5.3), and mechanisms for undertaking the needed work (Section 5.4).

5.1 Opinions about Another Coordinated Multiple-Contractor Project

One possible way to bridge the gaps in current technology might be another large scale coordinated project. If another large scale speech understanding project were undertaken, our respondents (Appendix 8, Section B-6.1) would favor development of several alternative systems, particularly if the systems address a spectrum of problem complexities, with one system directed at an easy problem, another at a moderately difficult task, and another at a quite difficult, challenging task. Other organizational features endorsed were the use of supporting research efforts conducted by specialist contractors, plus mid-term evaluations of the systems, extensive performance evaluation of the systems developed, and close interactions and frequent interchanges among contractors. The ARPA SUR project taught the value of mid-term milestones, back-up plans, and clearly demonstrated successes, rather than working on or modifying system features right up to the last minute.

The respondents were uncertain, or had mixed opinions, about completely defining fixed system specifications that must be achieved by fixed dates, and they questioned program management by committee. As one colleague said, "You can't legislate or schedule scientific breakthroughs".

When asked what system design choices they would make if a large scale follow-on project were undertaken, the respondents recommended: a moderate vocabulary of several hundred words or more; 10 to 100 speakers; three levels of system or language complexity; practical input through close talking microphones, telephones, or other communication channels of various qualities; systems adjustable to the speaker with only a few utterances; more substantial use of semantic and pragmatic constraints; near-real-time operation; and accuracy of 95-99%, to be achieved within a project period of three to five years. The primary differences from the ARPA SUR project are: a series of progressively more difficult tasks; more attention to practical needs like realistic input channels, many speakers, and high accuracy in real time; and no programmed demand for success on a fixed deadline.

5.2 The Need for Coordinated Studies and Computer Network Interactions

While such opinions may be of general interest, they do not necessarily demand the idea of another large-scale speech understanding project. Admittedly, for each of the problem areas listed in Figure 4-1 or in the respondent's list of gaps presented at the beginning of Section 4.4, separate programs could be undertaken. Yet, coordinated programs should obviously be defined that permit, and indeed promote, interactions and cooperation among various groups of researchers or developers. Many of these specific topics are intertwined, with the methods taken in one aspect affecting the best way to accomplish other tasks. A large scale, multiple-contractor research

program of cooperative and competitite developments of alternative systems with common goals does foster valuable interactions. Another mechanism for stimulating interactions and cooperation could be the recently-established voice interactions Technical Advisory Group (or "TAG) set up among United States Governmental funders of speech recognition work.

The ARPA SUR project was characterized by extensive interactions that were valuable to the successful progress of the research and systems developments. Several workshops were arranged, and regular communications were necessary among the members of the Steering Committee and among the various contractors. The ARPANET was an essential ingredient in all these interactions. "Mail" was communicated rapidly by ARPANET, thus avoiding postal delays and permitting rapid two-way interactions. Information could be rapidly distributed to many recipients. It is difficult to measure how much the project would have suffered without interactions over the ARPANET, but the impact would have been severe.

The ARPANET permitted support contractors to remotely use the large computer facilities of the system builders, thus avoiding duplicative efforts, travel costs, or cumbersome mailing of tapes and computer data. For example, SCRL made extensive use of CMU facilities over the ARPANET, for developing and using phonological analysis procedures. BBN provided Sperry Univac's prosodic analysis programs over the ARPANET. Speech segmentation files were transferred to CMU from various organizations for use in the 1973 CMU Speech Segmentation Workshop. The various segmentations could not have been compiled together and supplied to the participants in time for the workshop if it had not been for the ARPANET. The HEARSAY II system was developed with the SAIL language, which was developed and maintained (for CMU and other users) by Stanford Universtiy, with regular interactions over the ARPANET. Many other transfers of programs and data throughout the project proved essential to the progress made.

A prime example of the essential contributions of the ARPANET was the cooperative system development between SRI and SDC. Major programs were transferred from SRI to the SDC facility, and debugging and use of the programs were thus possible over the hundreds of miles between the facilities, without the expense and delays involved in otherwise-necessary travel, per diem, and splitting of the research teams.

Future projects can profit substantially from cooperative interactions such as the ARPA SUR project so effectively exemplified, and the ARPANET made possible. If another large-scale project were undertaken, or if any well-coördinated set of programs is attempted, computer network interactions and cooperative efforts in system development seem essential.

5.3 Specific Programs to be Undertaken

At least four types of speech recognition programs are needed now, including: 1. some applications studies with available commercial recognizers; 2. some comparative evaluations of alternative devices and specific improvements to handle noise, a large talker population, and limited forms of connected speech; 3. some advanced development projects to substantially expand recognition capabilities; and 4. research on necessary knowledge sources and basic concepts relevant to future success in recognition.

5.3.1 Applications Studies

We will discuss some military application projects in Section 6. Commercial applications studies (including human factor studies in actual field applications, such as were suggested in Section 4.5) are also needed, but will presumably be funded and conducted by commercial sources of recognizers.

5.3.2 Comparative Evaluations and System Improvements

Table 5.1 lists some of the projects needed to comparatively evaluate recognizers and extend their current capabilities. While a few studies have been done to compare speech input with other modalities of communication with a computer (e.g., Ochsman and Chapanis, 1974; Welch, 1977), more studies are still needed, particularly with realistic situations using actual computer input devices. Also, studies are needed to determine the effectiveness of human-to-computer communications under various constraints such as various vocabulary sizes, word confusabilities, structural varieties or complexities in the interactive language, "habitability" conditions, etc.

TABLE 5-1. NEEDED PROGRAMS FOR EVALUATING AND ADVANCING CURRENT TECHNOLOGY

COMPARATIVE EVALUATIONS OF ALTERNATIVE INPUT MODALITIES	Human-to-human simulation studies Input via available devices for each modality Effects of linguistic constraints on effectiveness of communications
COMPARATIVE EVALUATIONS OF ALTERNATIVE SPEECH RECOGNIZERS	Selection of databases and benchmark tasks Defining measures of task complexity Experimental evaluations under practical conditions
HUMAN FACTORS STUDIES	Effects of physical and mental stress Design criteria for user acceptance
REALISTIC CHANNEL CONDITIONS	Microphone characteristics and placement Telephone input Noise Channel distortions
EXTENSIONS OF RECOGNIZER CAPABILITIES WITHOUT MAJOR SYSTEM REDESIGN	Larger vocabularies Word spotting Speaker independence with extensive training

With the advent of many commerical sources of word recognizers has come a growing demand for objective procedures for evaluating systems. Several manuafacturers have mentioned to us the need for industry standards, such as general speech databases of 100 or more talkers speaking many instances of the digits and other vocabularies. Also, recognizers need to be comparatively evaluated with realistic "benchmark" tasks of interaction between unsophisticated users and voice entry systems. In addition, to facilitate evaluations without always using the same task, it would be advantageous to

develop good measures of task complexities, so one can compare 90% accuracy on an easy task with 60% accuracy on a difficult one. Other aspects of recognizer evaluation relate to various human factors studies and the effects of physical and mental stress on performance of the recognizer and human user. Systematic studies should be done to determine what accuracy and other system features are needed for user acceptance of the recognizer.

Tests are needed with various microphone characteristcs and placements, telephone input, noisy environment, and various channel distortions.

Finally some straight-forward extensions of recognizer capabilities should be considered that do not require major system re-design or advanced development projects, so that larger vocabularies, spotting of words in context, and speaker independence without extensive re-training will be possible.

5.3.3 Advanced Development Projects

A number of programs are called for in developing advanced recognition systems, as shown in Table 5-2. Extensive work is needed to bridge the gap between currently available (and usable) accurate isolated word recognizers, and long range work on ambitious speech understanding systems. Studies should be done to determine when it is truly profitable to use connected speech, and if so, how much language versatility is really useful for each application. Digit string recognizers and word sequence recognizers need to be refined substantially so that they provide accuracies comparable to those of word recognizers.

Speaker independence without the need for extensive training must also be assured. Similarly, programs are needed to develop practical speech understanding systems of capabilities exceeding those of Harpy, with HWIM and HEARSAY features effectively combined as outlined in Section A-1.2.3 of Appendix A. Comparison studies with task-independent continuous speech recognizers, such as the IBM system, still seem to be an option to restricted speech understanding systems, and prudence would suggest trying both options in future developments. For all types of continuous speech recognizers, work seems appropriate to assure fast processing so that systems can be quickly

TABLE 5-2. NEEDED PROGRAMS FOR DEVELOPING ADVANCED SYSTEMS

- EVALUATING THE NEED FOR CONTINUOUS SPEECH
- DIGIT STRING RECOGNIZERS
- WORD SEQUENCE RECOGNIZERS
- HARPY-LIKE LIMITED SPEECH UNDERSTANDING SYSTEMS
  - Applications to New Tasks
  - Refinements: Incremental compilation
Automatic knowledge acquisition
Additional knowledge sources
- MODERATELY-RESTRICTED SPEECH UNDERSTANDING SYSTEMS
  - (Combine HEARSAY and HWIM)
- TASK-INDEPENDENT CONTINUOUS SPEECH RECOGNIZERS
  - (IBM-like system)
- METHODS FOR FAST PROCESSING OF EXTENSIVE DATA

developed and tested with extensive speech data. Such speed considerations could be integral parts of other recognizer developments, but the importance of fast processing warrants explicit mention, and some work might even be appropriate on system-independent developments of fast processors.

Projects are already under way in recognition of digit strings and restricted word sequences at Bell Laboratories, Sperry Univac, Texas Instruments, Logicon, and ITT. The successful Harpy system and its integrated network structure are being adequately applied, extended, re-implemented on a smaller computer, and used in developing a system for computer-assisted training of air traffic controllers (Breaux, et al., 1977). Also, IBM (Bahl, et al., 1978) is developing a large continuous speech recognition system based on extensive statistics and a similar "generalized input-output" mathematical system structure (cf. Newell, 1975). It seems clear that such mathematically-based system structures will be getting adequate attention in future work, and we would caution system builders against confining their work to generalized input-output systems based on mathematical (statistical) analyses only, to the exclusion of knowledge sources based on phonetic, prosodic, and linguistic regularities. The primary gaps today appear to be in the aspects of multiple-knowledge source speech understanding systems, as shown in Table 5-3.

5.3.4 Research on Knowledge Sources and Recognition Concepts

In accordance with the opinions about priorities of "gaps" in speech understanding, which were assigned in Section 4.4, we list in Table 5-3 some of the top priority knowledge sources that need extensive further research and testing. Topping the list of priorities are studies of acoustic phonetic analysis, or the "front ends" of recognizers. Most acoustic parameters are adequate and effectively used, but there is need for some study of formant contours at transitions into and out of consonants. Some value might also be gained from the use of vocal tract area functions. New distance measures are needed that take into account the relative importance of various phonetic portions of speech. Improved segmentation and labeling is of high priority, particularly in what are currently the weaker aspects, such as establishing place of articulation of the consonants. Talker normalization procedures are needed to improve performance with a variety of talkers. One general aspect of speech science that has had little impact on recognition schemes and that deserves attention is the theory of speech perception.

The second major area in need of extensive research is in prosodic aids to recognition. Algorithms are needed for reliable detection of stresses, syllables, syntactic pauses, and intonational phrase boundaries, and schemes must be devised and tested for using prosodic information to aid phonemic analysis, word matching, and syntactic parsing. In addition, extensive controlled studies need to be done on basic prosodic correlates of linguistic structures. Prosodics research has been a weak area in comparison to segmental phonetic research, and extensive work is needed to bring it up to the level of maturity that has resulted from decades of intensive studies of segmental phonetics. Also, it may be fruitful to put the ideas about prosodic aids to recognition to the test, by developing and testing a total prosodically-guided speech understanding system.

Other high priority areas of research are listed in Table 5-3. We would like to reiterate the importance of system testing and performance evaluation. Methods are needed for comparatively evaluating two or more alternative systems.

TABLE 5-3. NEEDED PROGRAMS FOR RESEARCH ON NECESSARY CONCEPTS AND KNOWLEDGE SOURCES

GENERAL TOPICS	SPECIFIC TOPICS FOR RESEARCH
ACOUSTIC PHONETIC ANALYSIS	Use of vowel formant transitions Use of vocal tract area functions Improved distance measures Syntactic pattern recognition schemes Segmentation and labelling (especially distinguishing place of nasals, sibilants, fricatives, and stops) Talker normalization Study and use of perception models
PROSODIC AIDS TO RECOGNITION	Detecting stresses, syllables, syntactic pauses, and intonational phrase boundaries Prosodic aids to phonemic analysis and word matching Prosodic aids to parsing Research on prosodic correlates of linguistic structures Prosodically-guided speech understanding strategies
PERFORMANCE EVALUATION AND TASK COMPLEXITIES	Scoring procedures for selecting hypotheses Measuring complexities of tasks Relating task complexity to recognition scores Methods of system evaluation Evaluating system components and determining sources of errors
PHONOLOGICAL RULES	Coarticulation and acoustic phonetic rules Word boundary effects Generation of alternative word pronunciations
LINGUISTIC CONSTRAINTS ON AMBIGUITIES	Pragmatic constraints: discourse, task constraints Syntactic and semantic constraints Habitability

if they are not applied to tasks of equivalent complexity and similar or identical speech data. The beginning of an adequate evaluation of a system is to determine recognition accuracy on a large database of utterances, with known task complexity, but much more is needed to fully evaluate a system and its specific components and to determine the exact source of errors that occur.

Other work is needed on phonological rules, including coarticulation effects, detailed rules specifying acoustic parameter values for words, handling word boundary effects, and generating alternative word pronunciations from base forms in a lexicon. In the large-unit linguistic analysis of a system, work is needed on systematic methods for applying pragmatic constraints like discourse and task constraints on what could be said. Syntactic and semantic constraints should be considered so as to define a

hierarchy of useful languages of increasing complexities. The question of how habitable a language has to be to be useful and how current recognition languages compare with needed degree of habitability must be resolved. Other work must deal with scoring procedures for selecting among alternative hypotheses at various levels within a recognizer.

One important consideration for all recognition systems that deal with multiple knowledge sources is the way in which knowledge is represented. The field of artificial intelligence has played a major role in recent work within speech understanding system development. Its importance in the future cannot be underestimated and it is obvious that continued research in the field of knowledge understanding must be supported. Any system that employs multiple knowledge sources will have a need for the best representation of that knowledge.

5.4 Mechanisms for Undertaking Needed Work

We have listed a variety of applications (cf. Section 6), technical evaluations and refinements, system development efforts, and research topics that need attention. We reiterate that while separate programs could be undertaken in each of these areas, coordinated programs should obviously be defined that permit, and indeed promote, interactions and cooperation among contractors. Many of these specific topics are intertwined, with the methods taken in one aspect affecting the best way to accomplish other tasks. Another large-scale project, or highly cooperative projects coordinated through a Technical Advisory Group, are reasonable mechanisms to assure a well-run research program of cooperative and competitive developments of alternative systems and knowledge sources.

Project design and management should definitely have progressive milestones, and back-up plans ("Plan A, Plan B, etc.") so that intermediate results can be demonstrated, and intermediate systems are not thrown away but rather serve to verify past progress and the potential for future advancements. Finally, a substantial commitment to system testing and performance evaluation should be intrinsic to every system development.

We also recommend the establishment of two or three "speech science centers" with speech and linguistic expertise, powerful computer facilities and computer network capabilities, and mechanisms for visiting researchers to use such facilities to advance their work on various aspects of recognition, and to incorporate their advancements into the resident recognition systems. Such speech science centers could act as clearinghouses for useful speech databases, reports, research results, phonological rules, scoring procedures, recognition algorithms, and system structures, and could always offer working recognizers as research tools and testbeds for further advances. The centers could offer seminars and workshops for acoustic phonetic analysis, prosodics, phonological rules, interspeaker differences, etc. If affiliated with universities, they could provide the training and on-the-job experience appropriate for excellent new scholars in the field. Other speech research and development programs could be undertaken at the centers, such as in speech synthesis, speech transmissions systems, clinical speech studies, and linguistic analysis. With stable funding from a variety of sources, and visiting scholars on sabbaticals, or brief collaborative interactions, plus excellent facilities, such speech science centers could provide the concentrations of excellence needed to bridge the many gaps in current speech recognition technology.

Given the lead time needed to complete research projects and transfer the research into systems, research should begin as soon as possible on improved acoustic phonetic analyses, phonological rules, prosodic analysis methods, and performance metrics and evaluation procedures. A coordinated program in the development of recognizers of several distinct capabilities could then be progressively undertaken.

6. DOD APPLICATIONS FOR SPEECH RECOGNITION

Another aspect of a complete assessment of the technology of speech recognition is to consider how to accomplish "technology transfer", or the process of converting promising results with prototype recognizers into useful, economical devices that improve interactions with computers in everyday field applications. It was beyond the scope of this study to explore new uses for recognizers, or to add significantly to knowledge about existing applications. Several previous studies have already compiled lists of potential DOD applications for speech recognition (Turn, et al., 1974; Beek, et al., 1977; Feuge and Geer, 1978), and a number of agencies have already demonstrated the applicability of voice input to various important situations. What we offer here is a brief review and assessment of some DOD applications.

Our survey of expert opinions (cf. Appendix B, Section B-5.4) showed that the best applications are considered to be in command and control, air traffic control, data retrieval, inventorying, and package sorting. Table 6-1 shows some military applications that seem appropriate for available word recognizers. The agencies shown in parentheses have already funded or expressed interest in the applications as listed. Rome Air Development Center and the Defense Mapping Agency have already demonstrated the value of voice input of heights and depths at coordinates on a terrain map or oceanographic map, but further work in actual field use is called for, and it looks like connected digit recognition might help this hands-and-eyes-busy application. Logicon and the Naval Training Equipment Center have developed an excellent application in the training of air traffic controllers for ground controlled approach (cf. Section 3.3). The skills being learned are primarily verbal, and controllers must learn not to deviate from agreed-upon short utterances. Automation of instructor and pilot simulation functions is possible when the trainee's utterances are automatically recognized and corrected by machine responses. There may be many other such verbal training applications, which could prove as appropriate for speech recognizers as package sorting, inspection, and machine control have proven to be in commercial applications.

Another area of growing interest is cockpit communications, such as computer input aids for helicopter pilots (Coler, et al., 1978) and aircrews of P-3C aircraft. Boeing/Logicon defined a complete set of 26 speech recognition projects relevant to the P-3C aircrew and related applications (Feuge and Geer, 1978). The RADC efforts in having Texas Instruments develop an all-voice access control system for base internal security purposes is another application that deserves further work.

One promising but apparently untapped application for available recognizers is in computer assisted trouble shooting. An apprentice trouble shooter usually asks specific questions and communicates simple data to his more-knowledgeable supervisor, and such guided testing could be accomplished by interactions with a computer-stored trouble shooting information source.

Key word spotting has been the subject of considerable previous research and development projects, and still seems to be one of the simplest extensions of recognition capabilities to continuous speech. While its application to automatic surveilance of communication systems might be controversial, it can also serve as a word hypothesization process of a speech understanding system. Also, word spotting might be used in the captioning of television programs and training movies as an aid to the deaf.

TABLE 6-1. NEEDED PROGRAMS FOR APPLYING AVAILABLE RECOGNIZERS

- VOICE ENTRY OF CARTOGRAPHIC DATA (RADC, DMA)
- TRAINING AIR TRAFFIC CONTROLLERS (NTEC)
- COCKPIT COMMUNICATIONS
  - Helicopter pilots (NASA Ames)
  - Aircrew of P-3C aircraft (NADC, NTEC)
- ALL-VOICE ACCESS CONTROL TO SECURE AREAS COUPLED WITH SPEAKER VERIFICATION (RADC)
- COMPUTER-ASSISTED TROUBLE SHOOTING
- KEY-WORD SPOTTING (RADC,ONR)
  - Selecting conversations
  - Word hypothesization in a speech understanding system
  - Translation aids for the deaf

For each DOD application, studies can and should be done of the advantages and cost effectiveness of voice input facilities, the conditions and system specifications for the most appropriate form of recognizer, and the human factors of an effective system for interactions. The general analyses represented in the progressive steps of analysis in this report (i.e., defining relevant advantages and disadvantages, evaluating previous work, selecting the best current methods, and recommending programs to fill any remaining "gaps" in the necessary technology) can be applied to define specific programs for developing an adequate plan for development of the most appropriate speech recognition facility. For example, SCRL is under contract with NADC to apply the results of this general review of the field to the specific tasks of determining primary issues and techniques appropriate to speech recognition in cockpit communications. Similar studies could be useful so as to directly apply current speech recognition capabilities to practical DOD problems.

In essence, our review of the ARPA SUR project and survey of current technology has provided many specific recommendations that are appropriate for advancing speech recognition technology along directions that are needed for various DOD applications. As we illustrated in Section 1, there are many reasons for voice input to machines, though different applications have different needs. In addition to the well established technology in isolated word recognition (reviewed in Sections 1 and 3), we now have a developing spectrum of recognizers ranging from an imminent commercial

capacity in digit string recognition to more limited feasibility demonstrations of continuous word-sequence recognizers and restricted speech understanding systems. The ARPA SUR project successfully demonstrated (cf. Section 2 of Appendix A) that carefully constrained sentence understanding is possible and that several system structures and improved analysis components show promise of providing a brighter future of more versatile communications with computers. Yet, we have seen (Section 4) that many "gaps" remain before the more powerful recognition capabilities will be developed to the point where they are applicable as "off-the-shelf hardware" for operational DOD purposes. Primary among the gaps are "front end" aspects of recognition, performance evaluation and practical application studies, and essential research about phonetic, prosodic, phonological, and higher-level linguistic structures of spoken sentences. The various gaps can be bridged by the types of programs outlined in Section 5, which reiterate the concept of coordinated research and development in all aspects of recognition, using powerful facilities that might best be located at science centers of necessary interdisciplinary expertise. The technology seems ready for judicious but vigorous transfer into such DOD applications as cartography, training of skilled communicators and computer users, some cockpit communications, general inventorying and hands-free data entry, and computer instruction by unskilled workers and busy commanders. Now the task is to take the best aspects of this advancing technology and the recommendations given here, and implement a program for further advances and effective technology transfer.

7. REFERENCES

Bahl, L.R., J.K. Baker, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer (1978), Automatic Recognition of Continuously Spoken Sentences from a Finite State Grammar, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978, 418-421. Also, Recognition of a Continuously Read Natural Corpus, 422-424.

Baker, J.M. (1975), A New Time-Domain Analysis of Human Speech and Other Complex Waveforms, Technical Report CMUCSD, Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, PA.

Barnett, J.A., M.I. Bernstein, R.A. Goodman, and I.M. Kameny (1979), The SDC Speech Understanding System, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ : Prentic-Hall, Chapter 12.

Baudrey, M. and B. Dupeyrat (1978), Utilisation de Methodes Syntaxiques et de Filtrage Logique en Reconnaissance de la Parole, Congres AFCET/IRIA, Reconnaissance des Formes et Traitement des Images, Paris, February 21-23, 1978.

Beek, B., E.P. Neuberg, and D.C. Hodge (1977), An Assessment of the Technology of Automatic Speech Recognition for Military Applications, IEEE Transactions Acoustics, Speech, and Signal Processing, ASSP-25, Number 4, 310-322.

Bernstein, M., et al. (1976), Interactive Systems Research: Final Report to the Director, Advanced Research Projects Agency, Systems Development Corporation Report TM-5246/006/00, under ARPA Contract DAHC 15-73-C-0080, Santa Monica, CA.

Bobrow, D.G. and D.H. Klatt (1968), A Limited Speech Recognition System, in Proceedings of the AFIPS Fall Joint Computer Conference, Vol. 33, Washington D.C.: Thompson Book Co., 305-318.

Breaux, R. (1978), Laboratory Demonstration Computer Speech Recognition in Training, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control System Application, (R. Breaux, M. Curran, and E.M. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, December 6-8, 1977.

Breaux, R., P.M. Curran, and E.M. Huff (1978), Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application, NASA Ames Research Center, Moffett Field, CA, December 4-6, 1977; revised for distribution, April, 1978.

Chapanis, A. (1975), Interactive Human Communication, Scientific American, Vol. 232, 36-42.

Chapanis, A., R.N. Parrish, R.B. Ochsman, and C.D. Weeks (1977), Studies in Interactive Communication: II. The Effects of Four Communication Modes on the Linguistic Performance of Teams during Cooperative Problem Solving, Human Factors, 19, No. 2, 101-126.

Chomsky, N. (1957), Syntactic Structures. The Hague; Mouton.

Coler, C.R., E.M. Huff, R.P. Plummer, and M.H. Hitchcock (1978), Automatic Speech Recognition Research at NASA-Ames Research Center, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 143-170.

Connolly, D.W. (1978), Voice Data Entry in Air Traffic Control, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 171-196.

Cooper, F.S. (1976), Acoustic Cues in Natural Speech: Their Nature and Potential Uses in Speech Recognition, Final Report on ONR Contracts N00014-67-A-0129-002 and N00014-76-C-0591,. Haskins Laboratories, New Haven, CT.

Curran, M. (1978), Voice Integrated Systems, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 123-137.

Davis, K.H., R. Biddulph, and J. Balashek (1952), Automatic Recognition of Spoken Digits, The Journal of the Acoustical Society of America, Vol. 24, 637-645.

Dean, H.H. (1953), Effective Communication. New York: Prentice-Hall.

De Mori, R., S. Rivoira, and A. Serra (1975), A Speech Understanding System with Learning Capability, in Proceedings of the 4th International Joint Conference Artificial Intelligence, Tbilisi, USSR.

Denes, P.B. (1960), Automatic Speech Recognition: Experiments with a Recognizer Using Linguistic Statistics, U.S. Air Force Contract Number AF 61 (514)-1176, Technical Note number 4.

Denes, P. and M.V. Mathews (1960) Spoken Digit Recognition Using Time-Frequency Patterns Matching, The Journal of the Acoustical Society of America, Vol. 32, 1450-1455.

Doddington, G. (1976), Personal Identity Verification Using Voice, presented at ELECTRO 76, Boston, Mass.

Doddington, G.R., R.E. Helms, and B.M. Hydrick (1976), Speaker Verification III, Final Technical Report, Air Force Contract F 30602-75-C-0085, Report Number RADC-TR-UI-713804-F, Texas Instruments, Inc.

Dreyfus-Graf, J. (1949), Sonograph and Sound Mechanics, The Journal of the Acoustical Society of America, Vol. 22, 731-739.

Dudley, H., and S. Balashek (1958), Automatic Recognition of Phonetic Patterns in Speech, The Journal of the Acoustical Society of America, Vol. 30, 721-739.

Enea, H., and J. Reykjalin (1978), Low Cost Speech Recognition for the Personal Computer Market, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 285-288.

Erman, L.D. and V.R. Lesser (1979), The HEARSAY-II Speech Understanding System: A Tutorial, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, N.J.: Prentice-Hall, Chapter 16.

Fejfar, A. (1977), Test Results: Advanced Development Models of BISS Identity Verification Equipment (Volume 1, Executive Summary), MITRE Technical Report MTR-3442, The MITRE Corporation, Bedford, MA.

Feuge, R.L, and C.W. Geer (1978), Integrated Applications of Automated Speech Technology, Final Report and Program Plan on ONR Contract N00014-77-C-0401, Boeing Aerospace Company (with Logicon, Inc.), Seattle, WA.

Flanagan, J.L., S. Levinson, L.R. Rabiner, and A.E. Rosenberg (1979), Techniques for Expanding the Capabilities of Practical Speech Recognizers, in Trends in Speech Recognition (W.A. Lea, Editor), Englewood Cliffs, NJ: Prentice-Hall, Chapter 18.

Focht, L.R. (1963), The Single Equivalent Formant, IEEE International Communication Conference Digest, Philadelphia, PA, 108-115.

Forgie, J.W., (1974), Speech Understanding Systems-Semiannual Technical Summary Report, M.I.T. Lincoln Laboratory, Lexington, MA.

Forgie, J.W. and C.D. Forgie (1959), Results Obtained from a Vowel Recognition Computer Program, The Journal of the Acoustical Society of America, Vol. 31, 1480-1489.

Forgie, J.W. and C.D. Forgie (1961), Automatic Method of Plosive Identification, The Journal of the Acoustical Society of America, Vol. 34, 1979(A).

Hanson, B., D. Brill, M. Earle, E. Hayden, M. Mines, H. Neu, B. Oshika, and J. Shoup-Hummel (1976), Techniques for Natural Speech Processing, ARPA SUR Note 209, Final Report on Contract N00014-73-C-0221, Speech Communications Research Laboratory, Santa Barbara, CA.

Fry, D.B. and P. Denes (1958), The Solution of Some Fundamental Problems in Mechanical Speech Recognition, Language and Speech, 1, 35-38.

Fu, K.S. (1974), Syntactic Methods in Pattern Recognition. New York: Academic Press.

Gillman, R.A. (1975), A Fast Frequency Domain Pitch Algorithm, Journal of Acoustical Society of America, Vol. 58, Suppl., Fall, 1975, S62 (A).

Goldberg, H.G. (1975), Segmentation and Labeling of Speech: A Comparative Performance Evaluation, Technical Report CMUCSD, Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, PA.

Goodman, G. (1976), Analysis of Languages for Man-Machine Voice Communication, Technical Report CMUCSD, Ph.D. Dissertation (through Stanford University), Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Goodman, G., D. Scelza, and B. Beek (1977), An Application of Connected Speech to the Cartography Task, Proceedings of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, 811-814.

Grady, M.W., M.B. Hicklin, and J.E. Porter (1978), Practical Applications of Interactive Voice Technologies--Some Accomplishments and Prospects, Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command/Control Systems Application (R. Breaux, M. Curran, and E. Huff, Editors), NASA Ames Research Center, Moffett Field, CA, 217-233.

Halle, M. and K.N. Stevens (1962), Speech Recognition: A Model and a Program for Research, IRE Transactions on Information Theory, IT-8, 155-159.

Hanson, B., et al. (1976), Techniques for Natural Speech Processing, ARPA SUR Note 209, Final Report on Contract N00014-73-C-0221, Speech Communications Research Laboratory, Santa Barbara, CA.

Haton, J.P. (1979), Speech Recognition Work in Western Europe, in Trends in Speech Recognition, (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 24.

Hemdal, J.F. and G.W. Hughes (1967), A Feature Based Computer Recognition Program for the Modeling of Vowel Perception, in Models for the Perception of Speech and Visual Form, W. Wathen-Dunn. Editor, Cambridge, MA: M.I.T. Press.

Hill, D.R. (1971), Man-Machine Interaction Using Speech, in Advances in Computers (F.L. Alt, M. Rubinoff, and M.C. Yovits, Editors). New York: Academic Press, 11, 165-230.

Itakura, F. (1975), Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Transactions on Acoustics Speech, and Signal Processing, Vol. ASSP-23, 67-72.

Jassem, W. (1979), Speech Recognition Work in Poland, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 23.

Jelinek, F. (1976), Continuous Speech Recognition by Statistical Methods, Proceedings of the IEEE, Volume 64, Number 4, 532-556.

Kelley, T.P., J.T. Martin, and J.R. Barger (1968), Voice Controller for Astronaut Manuevering Unit, Technical Report AFAL-TR-68-308, Air Force Avionics Laboratory, Wright Patterson Air Force Base, OH.

Klatt, D.H. (1976), A Digital Filter Bank for Spectral Matching, Proceedings of the 1976 IEEE International Conference on Acoustics Speech and Signal Processing, Philadelphia, PA. (IEEE Catalog No. 76H1067-8 ASSP), 537-540.

Klatt, D.H. (1977), Review of the ARPA Speech Understanding Project, Journal of the Acoustical Society of America, 62 , 1345-1366.

Klatt, D.H. (1979a), Overview of the ARPA Speech Understanding Project, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 11.

Klatt, D.H. (1979b), Scriber and LAFS: Two New Approaches to Speech Analysis, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 25.

Klatt, D.H. and K.N. Stevens (1972), Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching, Proceedings 1972 Conference on Speech Communication and Processing, IEEE and AFCRL: Bedford, MA, 315-318.

Kuhn, G.M. (1975), On the Front Cavity Resonance and Its Possible Role in Speech Perception, The Journal of the Acoustical Society of America, Vol. 58, 578-585.

Lea, W.A. (1966), The 'Spectrum' of Weak Generative Powers of Grammars, Mechanical Translation, 9, 10-14.

Lea, W.A. (1969), The Impact of Speech Communication with Computers, Proceedings of the Sixth Space Congress, Vol. 1, Cocoa Beach, FL; Brevard Printers, March, 1969, pp. 15-19 to 15-31.

Lea, W.A. (1970a), Towards Versatile Speech Communication with Computers, International Journal of Man-Machine Studies, 2 , 107-155.

Lea, W.A. (1970b), Evaluating Speech Recognition Work, Journal of the Acoustical Society of America, 47, 1612-1614.

Lea, W.A. (1972), Computer Recognition of Speech, Current Trends in Linguistics (Vol. 12: Linguistics and Adjacent Arts and Sciences). The Hague: Mouton, 1561-1620.

Lea, W.A. (1973a), An Approach to Syntactic Recognition Without Phonemics, IEEE Transactions on Audio and Electroacoustics, AU-21, 249-358.

Lea, W.A. (1973b), An Algorithm for Locating Stressed Syllables in Continuous Speech, Journal of the Acoustical Society of America, 55, 411(A).

Lea, W.A. (1973c), Evidence that Stressed Syllables Are the Most Readily Decoded Portions of Continuous Speech, Journal of the Acoustical Society of America, 55, 410(A).

Lea, W.A. (1974), Prosodic Aids to Speech Recognition: IV. A General Strategy for Prosodically-Guided Speech Understanding, Univac Report No. PX10791. Sperry Univac, DSD, St. Paul, MN.

Lea, W.A. (1976), Prosodic Aids to Speech Recognition: IX. Acoustic-Prosodic Patterns in Selected English Phrase Structures, Univac Report No. PX11963, Sperry Univac DSD, St. Paul, MN.

Lea, W.A. (1977) Contributions of Speech Science to the Technology of Man-Machine Voice Interactions, Proceedings of the Workshop in Voice Technology for Interactive Real-Time Command/Control System Application (R. Breaus, M. Curran, and E.M. Huff, Editors) NASA Ames Research Center, Moffett Field, CA, December 6-8, 1977.

Lea, W.A. (1978), Voice Input to Computers: A Critical Overview, invited presentation to the National Computer Conference, Anaheim, CA, June 8, 1978.

Lea, W.A. (1979a), Editor: Trends in Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall.

Lea, W.A. (1979b), The Value of Speech Recognition Systems, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 1.

Lea, W.A. (1979c), Speech Recognition: Past, Present, and Future, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 4.

Lea, W.A. (1979d), Prosodic Aids to Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 8.

Lea, W.A. (1979e), Speech Recognition: What is Needed Now?, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 26.

Lea, W.A. and D.R. Kloker (1975), Prosodic Aids to Speech Recognition: VI. Timing Cues to Linguistic Structures, Univac Report No. PX11534, Sperry Univac DSD, St. Paul, MN.

Lea, W.A., M.F. Medress, and T.E. Skinner (1975), A Prosodically-Guided Speech Understanding Strategy, IEEE Transactions in Acoustics, Speech, and Signal Processing, ASSP-23, 30-38.

Lea, W.A. and J.E. Shoup (1977), Specific Contributions of the ARPA SUR Project to Speech Science, The Journal of the Acoustical Society of America, Vol. 62, Supplement 1.

Lea, W.A. and J.E. Shoup (1978a), Gaps in the Technology of Speech Understanding, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processings, Tulsa, OK, (IEEE Catalog No. 78CH1285-6 ASSP), 405-408.

Lea, W.A. and J.E. Shoup (1978b), Recommendations for Advancing Speech Recognition, Journal of the Acoustical Society of America, 63: Supplement 1, S78(A).

Lea, W.A. and J.E. Shoup (1979), Specific Contributions of the ARPA SUR Project, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 17.

Lindgren, N. (1965), Machine Recognition of Human Language, IEEE Spectrum, Vol. 2, March, April, May issues.

Lowerre, B.T. (1976), The Harpy Speech Recognition System, Technical Report CMUCSD, Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, PA.

Lowerre, B.T. and D.R. Reddy (1979), The Harpy Speech Understanding System, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 15.

Makhoul, J. and C.C. Cook (1974), Optimal Number of Poles in a Linear Prediction Model, Journal of the Acoustical Society of America, 56, S14(A).

Mariani, J.J. and J.S. Lienard (1978), ESOPE Ø: un programme de comprêhension automatique de la parole procédant par prédiction-vêrification aux niveaux phonétique, lexical et syntaxique, Congres AFCET/IRIA Reconnaissance des Formes et Traitment des Images, Paris, France.

Martin, T.B. (1976), Practical Applications of Voice Input to Machines, Proceedings of the IEEE, Volume 64, Number 4, 487-501.

Martin, T.B. and J. Welch (1979), Practical Speech Recognizers and Some Performance Evaluation Parameters, Trends in Speech Recognition, W.A. Lea, Editor, Englewood Cliffs, NJ: Prentice-Hall, Chapter 3.

McCandless, S.S. (1974), Use of Formant Motion in Speech Recognition, Proceedings of the IEEE Symposium on Speech Recognition (IEEE Catalog No. 74H0878-9AE), Carnegie-Mellon University, Pittsburgh, PA, 211.

Medress, M.F. (1972), A Procedure for Machine Recognition of Speech, Conference Record of the 1972 Conference on Speech Communication and Processing. Newton, MA: IEEE Catalog Number AD-742236, 113-116.

Medress, M.F. (1979), The Sperry Univac System for Continuous Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 19.

Medress, M.F., T.C. Diller, D.R. Kloker, L.L. Lutton, H.N. Oredson, and T.E. Skinner (1978), An Automatic Word Spotting System for Conversational Speech, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, IEEE Catalog Number 77CI1197-3 ASSP, 468-473.

Medress, M.F., T.E. Skinner, D.R. Kloker, T.C. Diller, and W.A. Lea (1977), A System for the Recognition of Spoken Connected Word Sequences, Proceedings of the 1977 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, IEEE Catalog No. 77CH1197-3ASSP, 468-473.

Mermelstein, P. (1975), Automatic Segmentation of Speech into Syllabic Units, The Journal of the Acoustical Society of America, 58, 880-883.

Mermelstein, P. (1976), Distance Measures for Speech Recognition: Psychological and Instrumental, in Pattern Recognition and Artificial Intelligence. New York: Academic Press, 374-387.

Moshier, S.L., P.N. Leiby, and R.E. Smith (1977), Key Word Classification, Air Force Report Number RADC-TR-77-122, Final Report on Air Force Contract Number F30602-75-C-0171, Rome Air Development Center, Griffiss AFB, NY.

Nakagawa, S.I. (1976), A Machine Understanding System for Spoken Japanese Sentences, Ph.D. Dissertation, Department of Information Sciences, Kyoto University, Kyoto, Japan.

Neuburg, E.P. (1975), Philosophies of Speech Recognition, Speech Recognition: Invited Papers of the 1974 IEEE Symposium (D.R. Reddy, Editor), New York: Academic Press, 83-95.

Neuburg, E.P. (1979), Needs versus Competence in Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 2.

Newell, A. (1975), A Tutorial on Speech Understanding Systems, Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium (D.R. Reddy, Editor), New York: Academic Press, 3-54.

Newell, A. (1978), Harpy, Production Systems, and Human Cognition, unpublished report, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Newell, A., J. Barnett, J. Forgie, C. Green, D.H. Klatt, J.C.R. Licklider, J. Munson, D.R. Reddy, and W.A. Woods (1971), Speech Understanding Systems: Final Report on a Study Group, Carnegie-Mellon University, Pittsburgh, PA. (Reprinted by American Elsevier, Amsterdam, North-Holland, 1973).

Niimi, Y., Y. Kobayashi, T. Asami, and Y. Miki (1975), The Speech Recognition System of 'SPOKEN-BASIC', Proceedings of the 2nd USA-JAPAN Computer Conference, Tokyo, Japan, 375.

Nye, J.M. (1979), The Expanding Market for Commercial Speech Recognizers, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 20.

Ochsman, R.B. and A. Chapanis (1974), The Effects of 10 Communication Modes on the Behaviour of Teams During Co-Operative Problem-Solving, International Journal Man-Machine Studies, Volume 6, 579-619.

Peterson, G.E. (1961), Automatic Speech Recognition Procedures, Language and Speech, 4, 200-219.

Pierce, J.R. (1969), Whither Speech Recognition?, Journal of the Acoustical Society of America, 46, 1049-1051.

Pratt, A.W., A.H. Roberts, and K. Lewis (1966), Seminar on Computational Linguistics, National Institutes of Health, Bethesda, MD. Public Health Service Publicaiton No. 1716. Washington, DC: Superintendent of Documents.

Reddy, D.R. (1973), The CMU Speech Understanding Project-Progress Report, October 15, 1973, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Reddy, D.R., (1976), Speech Recognition by Machine: A Review, Proceedings of the IEEE, Vol. 64, 501-531.

Reddy, D.R., et al. (1976), Speech Understanding Systems: Summary of Results of the Five Year Research Effort at Carnegie-Mellon University (2nd version, 1977), Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa.

Rosenberg, A.E. and C.E. Schmidt (1977), Recognition of Spoken Spelled Names Applied to Directory Assistance, The Journal of the Acoustical Society of America, Vol. 62, Supplement 1, S63.

Sambur, M.R. and L.R. Rabiner (1976), A Statistical Decision Approach to the Recogniiton of Connected Digits, IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume ASSP-24, Number 6, 550-558.

Scott, P.B. (1976), Voice Input Code Identifier, Final Technology Report, Air Force Contract F30602-75-C-0111, Report Number RADC-TR-77-190, Rome Air Development Center, Air Force Systems Command, Grifffiss AFB, NY.

Sekiguchi, Y. and M. Shigenaga (1978), Speech Recognition System for Japanese Sentences, Journal of the Acoustical Society of Japan, Volume 34, Number 3 204-213.

Schwartz, R.H. (1976), Acoustic-Phonetic Experiment Facility for the Study of Continous Speech, Proceedings of the 1978 International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 1-4.

Schwartz, R.H. and V.W. Zue (1976), Acoustic-Phonetic Recognition in BBN SPEECHLIS, Proceedings of the 1976 International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 21-24.

Shannon, C.E. and W. Weaver (1949), The Mathematical Theory of Communication. Reprinted 1962, University of Illinois Press, Urbana.

Shockey, L. and D.R. Reddy (1974), Quantitative Analysis of Speech Perception: Results from Transcription of Connected Speech from Unfamiliar Languages, Speech Communication Seminar, Stockholm, Sweden.

Shoup, J.E. (1979), Phonological Aspects of Speech Recognition, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 6.

Skinner, T.E. (1973), Speech Parameter Extraction: Fundamental Frequency, Spectral, and Formant Frequency Processing, Univac Report No. PX10376, Univac Park, St. Paul, MN.

Smith, A.R. and M.R. Sambur (1979), Hypothesizing and Verifying Words for Speech Production, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentic-Hall, Chapter 7.

Sondhi, M.M. (1968), New Methods of Pitch Extraction, IEEE Transactions on Audio and Electroacoustics, AU-16, 262-266.

Sondhi, M.M. and S.E. Levinson (1977), Relative Difficulty and Robustness of Speech Recogniiton Tasks that Use Grammatical Constraints, Journal of the Acoustical Society of America, 63, Supplement 1, S64(A).

Sondhi, M.M. and S.E. Levinson (1978), Computing Relative Redundancy to Measure Grammatical Constraint in Speech Recognition Tasks, Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK.

Teacher, C.F., H.G. Kellett, and L.R. Focht (1967), Experimental Limited Vocabulary Speech Recognizer, IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, 127-130.

Turn, R., A. Hoffman, and T. Lippiatt (1974), Military Applications of Speech Understanding Systems, Report Number R-1434-ARPA, Rand Corporation.

Vicens, P.J. (1969), Aspects of Speech Recognition by Computer. Technical Report, Stanford University, AI Memo 85, Stanford, CA, (Ph.D. Dissertation).

Wakita, H. and S. Makino (1979), Speech Recognition Work in Japan, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 22.

Walker, D.E. (1979), SRI Research on Speech Understanding, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 13.

Walker, D.E., R. Fikes, B. Grosz, G. Hendrix, W. Paxton, A. Robinson, J. Robinson, and J. Slocum (1976), Speech Understanding Research, Final Technical Report, 15 October 1975 to 14 October 1976, Stanford Research Institute, Menlo Park, CA.

Walker, D.E. (1977), Speech Understanding and AI; AI and Speech Understanding, panel discussion, Proceedings of the Fifth International Joint Conference on Artificial Intelligence, MIT, Cambridge, MA, August 22-25, 1977, Vol. Two, 970-974.

Weinstein, C.J., S.S. McCandless, L.F. Mondshein, and V.W. Zue (1975), A System for Acoustic-Phonetic Analysis of Continuous Speech, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 54-67.

Welch, J.R. (1977), Automatic Data Entry Analysis, Final Technical Report RADC TR-77-306, Rome Air Development Center, Rome, NY.

White, G.M. (1978), Continuous Speech Recognition, Dynamic Programming, Knowledge Nets and Harpy, Proceedings of Wescon, Los Angeles, CA, September 12-14, 1978, 28/2.

White, G.M. and R.B. Neely (1975), Speech Recognition Experiments With Linear Prediction, Bandpass Filtering, and Dynamic Programming, IEEE Transactions Acoustics, Speech, and Signal Processing, Volume ASSP-24, 183-188.

White, G.M. and M.R. Sambur (1979), Speech Recogntion Research at ITT Defense Communications Division, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 21.

Willems, Y. (1972), The Use of Prosodics in the Automatic Recognition of Spoken English Words, Ph.D. Dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Wiren, J. and H.L. Stubbs (1956), Electronic Binary Selection System for Phoneme Classification, The Journal of the Acoustical Society of America, Vol. 28, 1082-1091.

Wolf, J.J. (1976), Speech Recognition and Understanding, in Pattern Recognition, K.S. Fu, Editor, New York: Springer.

Wolf, J.J. and W.A. Woods (1979), The HWIM Speech Understanding System, Trends in Speech Recognition (W.A. Lea, Editor). Englewood Cliffs, NJ: Prentice-Hall, Chapter 14.

Woods, W.A. (1975), Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 2-10.

Woods, W.A. et al. (1976), Speech Understanding Systems, BBN Report No. 34-38 (5 Volumes), Final Report on ONR Contract No. N00014-75-C-0533, Bolt, Berenek and Newman, Cambridge, MA.

Yilmaz, H. (1967), A Theory of Speech Perception, Bulletin of Mathematical Biophysics, Vol. 29.

APPENDICES

APPENDIX A:
DETAILED CONTRIBUTIONS
OF THE
ARPA SUR PROJECT

While it is quite appropriate to evaluate the contributions of the ARPA SUR project in the light of its initial goals, as was done in Section 2 of this report, it is also useful to ask in what ways the project has contributed to the total context of speech recognition work and other research and development work in speech science, computer science, and artificial intelligence. We consider the system contributions, including the Harpy system structure (Sec. A-1.1), the experimental testing of alternative structures and control strategies (Sec. A-1.2), and the methods and results in performance evaluation (Sec. A-1.3). The detailed contributions in specific aspects of recognition techniques will be summarized in Section A-2.

A-1. DETAILED CONTRIBUTIONS IN SYSTEM DESIGNS

A-1.1 The Harpy System

Harpy has received considerable attention because it is one of the first systems to attain high performance in large vocabulary continuous speech recognition, and its contributions include its public demonstration of the credibility of task-constrained continuous speech input to computers. Ultimate practicality is suggested by the facts that this was accomplished at an order of magnitude faster than initially called for, with half the allowable error rate, under practical conditions of somewhat noisy speech tranduced by an inexpensive close talking microphone. As one developer of a competing system said to us, no one can now reasonably develop speech understanding systems without understanding Harpy, and using it as a "benchmark" system for comparisons with new recognizers.

To achieve success, Harpy used two primary system design contributions: the integrated network representation of knowledge and the beam search technique. The integrated network structure collapses knowledge at various levels (phonetic, phonological, lexical, and syntactic) into one generative model of acceptable pronunciations for recognizable sentences. As described in detail by Lowerre and Reddy (1979), Harpy begins with a word network of acceptable word sequences, in which the nodes are words and any path through the network gives an acceptable sentence. Then each word node gets replaced by a pronunciation network, representing expected pronunciations of the word. Word boundary rules, initially entered by hand, operate on the network to handle phone string variations due to influences of each word on its neighbors. During the automatic compiling of a single composite network, optimization heuristics are used to yield an efficient network of correct pronunciations. The network for the "document retrieval" demonstration task had almost 15,000 nodes, and required 13 hours of PDP-10 computer time to develop. The nodes in this fully-expanded pronunciation network are "phones" or allophonic segments taken from a vocabulary of 98 alternative spectral templates.

Harpy finds the best match between each of the incoming acoustically-derived segments and the 98 phonetic templates, using Itakura's (1975) distance metric for scoring matches. At each segment in the left-to-right search, the best match and some near misses are retained for further testing. The number of near misses (or "beam width") retained at each point is dynamically adjusted as the search progresses through the network.

The finite state "network" representation (or "Markov model") is not new, and, in 1957, Chomsky argued that if one accepts no fixed upper limit on the complexity of English sentences, then English (and certain major subsets of English) could not be represented by such a finite state graph (cf. Lea, 1966). Indeed, unless the language initially designed for Harpy is in the restricted class of such "finite state languages", the network that does include all the acceptable pronunciations will necessarily also allow pronunciations that were <u>not</u> in the intended language. This is part of the reason why Wolf and Woods (1979) question the extendibility of Harpy to other useful tasks. Harpy acts like an overgrown "word verifier", matching expected pronunciations of the <u>total</u> utterance to the total incoming sequence of short time segments. It is a natural "next step" from previous isolated-word recognizers. In fact, it is basically a heuristic form of dynamic programming, which has become very popular in isolated word recognizers (White, 1979).

Dragon and IBM recognition systems have also used Markov models or finite state networks. What Harpy's developers did was to successfully combine the best features of previous systems such as Dragon's network representation and HEARSAY I's phonetic segmentation procedures. The beam search avoided the expensive, time consuming processes of an "admissible" recognition strategy (which Dragon used and HWIM was intended to use, to <u>guarantee</u> finding the optimal solution in the search for the best match with the input data). The network and beam search also achieved most of the delayed-decision advantages of HEARSAY I's "best-first" strategy, by pursuing how an apparently promising phonetic-sequence hypothesis fits at lexical and syntactic levels before pursuing less promising alternatives. (SRI also found that it is good to use information from several knowledge sources to avoid erroneous interpretations and reinforce choices of each knowledge source.) Backtracking (after a high-scoring hypothesis fails) was effectively avoided by Harpy's beam search. Occasionally the correct interpretation will <u>not</u> be among the high-scoring alternatives in the beam, so that a failure to recognize may occur, but obviously that did not occur frequently enough to reduce accuracy extensively.

Some problems with Harpy are noted by Wolf and Woods (1979) and by Lowerre and Reddy (1979). Making a pronunciation dictionary for a new vocabulary is very time consuming, and it would be useful to have automatic methods for learning new word structures (and new syntactic constructions) directly from example pronunciations. Network compiling is very expensive in computer time. Juncture rules, pronunciation variabilities, and duration effects are not easily represented as a graph. The juncture rules had to be manually tailored to the task. Here again, automatic knowledge acquisition would be valuable. Harpy is very sensitive to <u>missing</u> states in the incoming phonetic sequence, unless extensive effort is devoted to creating optional states during the network creation time. In general, the collapsing of all types of information intoa single network makes it difficult to revise minor parts of

the incorporated knowledge, such as adding new words or new pronunciations to old words, incorporating newly learned word-juncture or phonological rules, adding new structures, introducing prosodic rules, etc.

An important contribution of Harpy was its demonstration of the value of using syntactic (and other linguistic) constraints to constrain the recognition problem, rather than just having the linguistic knowledge adequately model a large subset of English sentences. This was one of the original premises (or "dogmas"; Newell, 1975) of speech understanding that was vividly verified by Harpy's success.

A-1.2 Alternative Structures and Control Strategies

A-1.2.1 Several System Structures - There is a real danger that Harpy's success may overshadow other equally important contributions of the ARPA SUR project. Another major contribution was the variety of system structures and control strategies that were developed and applied for the first time to the recognition of continuous speech. HEARSAY II came very close to meeting the system specifications, using a generalized structure that is in some senses at the other extreme from Harpy; that is, HEARSAY II had clearly separated, cooperating knowledge sources that could be selectively modified, independently tested, and augmented by new knowledge sources without disrupting the total system. HEARSAY II's blackboard permitted an anonymity and independence of system modules such that the function and the very existence of each knowledge source was not necessary or crucial to the others, yet they could use and correct each other's hypotheses through the common blackboard. HWIM complemented HEARSAY II's fixed structure but readily variable components by having more-fixed components, and flexible control strategies. It was no small task to implement such multiple knowledge source systems that could handle the variabilities in continuous speech.

Woods (in Walker, et al., 1977, p. 972) has aptly noted that several of the ARPA SUR systems used a "factored knowledge structure", in which common parts of different knowledge sources are merged in such a way that retrieval processes can access them incrementally to progressively create more and more specific hypotheses about the utterance. Thus, Harpy merged grammar, lexical forms, rules, and pronunciations, while HWIM merged phonetic, phonological and lexical information in the phonetic lattice and lexical decoding network, and syntactic and semantic information in the ATN grammar. This trend seems to be a promising contribution.

For completeness, we should also note the performances of some intermediate systems developed in the course of the project, as shown in Fig. A-1. The earlier systems worked with more limited vocabularies and tasks, achieving various low-to-moderate levels of sentence understanding accuracy. The HEARSAY I system used heavy semantic constraints about the status of a chess board and the allowable and plausible moves, to do 79% correct understanding of spoken chess moves. However, it did much poorer on other less-constrained tasks. CMU researchers believe HEARSAY I could have achieved the ARPA SUR goals, given the acoustic-phonetic capabilities of the final (1976) systems (Reddy, et al., 1976, p. 4). One of the attractive features of the HEARSAY I system was its use of independent cooperating knowledge sources or modules, which could be removed one at a time to establish the performance with, versus without, that module. Such "ablation studies" (Newell, 1975) help determine the contribution of each system component, and permit detecting "weak links" in the system operation.

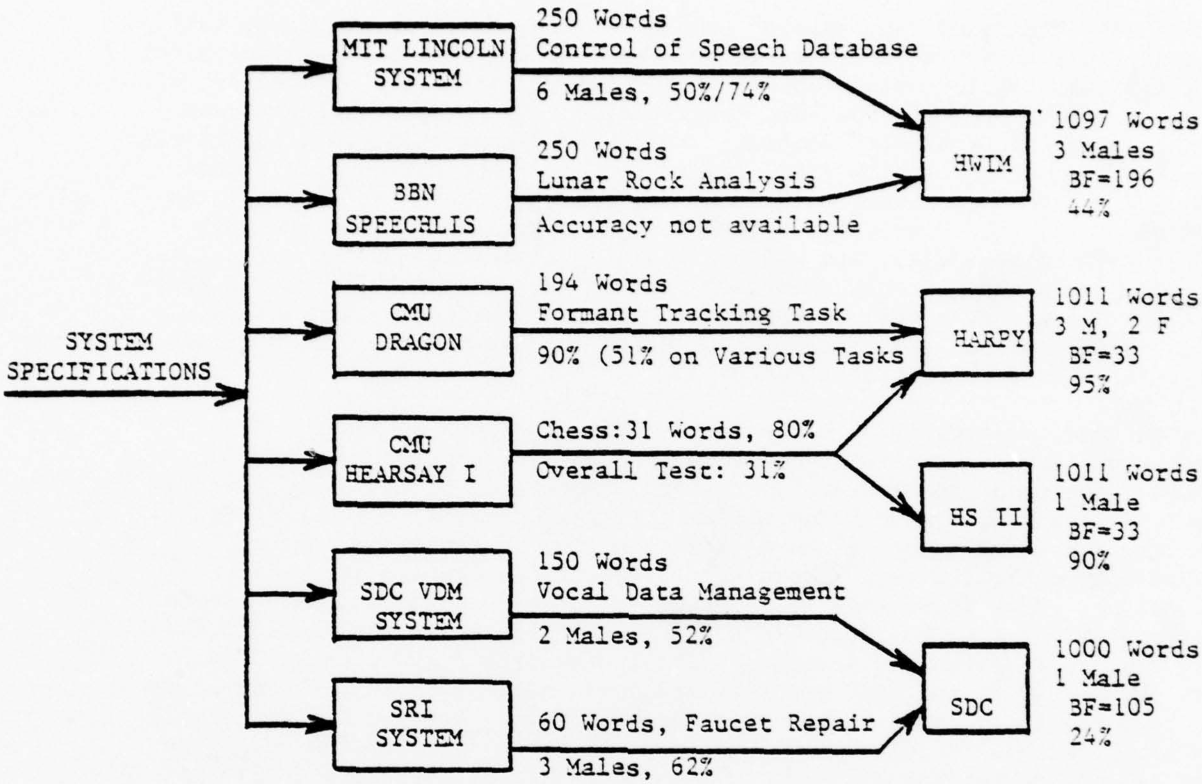


Fig. A-1. Performance results for systems developed throughout the ARPA SUR project. (Each task is briefly described, and accuracy percentages are given for correct semantic understanding.)

Lincoln Laboratories' system was one of the best performing intermediate systems, using a 250-word vocabulary and some of the best acoustic phonetic processing techniques available. Some of the "front-end" processing ideas of the Lincoln system were incorporated into the HWIM system and other systems (e.g., the Sperry Univac systems; Medress, et al., 1977). Dragon, the predecessor of Harpy, used a Markov model, no phonetic segmentation, and a highly constrained task to achieve fairly good performance. SDC's Vocal Data Management system had an intermediate performance of 52% correct recognition of sentences involving a vocabulary of 150 words. Also, SRI's system working with a small vocabulary of 60 words obtained 62% recognition at the midpoint of the ARPA SUR project.

A-1.2.2 Experimenting with Alternative Control Strategies - Besides building total systems with interesting structures, the ARPA SUR project also contributed extensive evidence about the advantages and disadvantages of various structures. BBN explored a number of different control strategies, including "left-to-right" versus "middle-out" analyses, and a "hybrid" strategy whereby analysis was initially anchored in the first stressed word in the utterance, then any extensions to the left (to the beginning of the utterance) were attempted before extensions to the right were attempted.

They sought efficient "admissible" control strategies which guarantee the best possible interpretation, and found one superior admissible method based on their "shortfall density" concept for scoring alternative hypotheses (cf. Wolf and Woods, 1979). However, for efficiency they also developed (and finally used) "approximate" methods. They showed that their best admissible strategy took only slightly (30%) longer than the best approximate method and was almost as accurate, with fewer misleading acceptances of wrong interpretations. Unfortunately, this promising admissible method was only tested with 10 sentences, and was not used in their final system demonstrations. Dragon and the HWIM shortfall density method used the first "admissible" strategies developed in the field, and such systematic approaches to devising control strategies were a marked improvement over the usual ad hoc developments of control structures.

SRI (cf. Walker, 1979) conducted experiments that evaluated sixteen alternative control strategies, and the large variations they obtained in accuracy and runtime confirm the importance of the control strategy in determining system performance. A primary conclusion was that it paid (in increased accuracy and reduced runtime) to check the context around a hypothesized word before setting priorities on all alternative hypotheses. They found that their forms of island driving and focusing on what are expected to be important areas of the speech did not help (in fact, they hindered) system performance. However, what this study as well as work on HEARSAY II and HWIM actually seems to indicate is that island driving is going to increase accuracy or reduce runtime only if (1) the islands are true "islands of reliability", such as might be assured with multiple-word islands or perhaps reliably-encoded stressed syllables; and (2) the expansion of islands by hypothesizing surrounding contexts can be controlled (such as by one-way expansions or BBN's hybrid strategy), so that combinatiorial explosions of alternative extensions do not arise.

A-1.2.3 Comparative Evaluations of the ARPA SUR Systems - Comparison of the ARPA SUR systems is a difficult but necessary part of the total assessment of the project's contributions. To begin with, we can exclude SDC's system from detailed consideration, since SDC researchers themselves consider that even a flawless version of the 1976 system would have improved performance only slightly, and it is not now a workable system suitable for further development without drastic changes. (They include in their list of important changes needed: the acoustic phonetic processor, the scoring algorithms, the phonological rules, the mapper, and the higher-level components such as semantics and discourse (Bernstein, et al., 1979). Little of the system would stay the same.)

Harpy's accuracy makes it appear to be the leading system contribution, but the counter-arguments given by HWIM and HEARSAY II developers (Wolf and Woods, 1979; Erman and Lesser, 1979) deserve attention, as do some frank admissions by Lowerre and Reddy (1979) of the difficulties of building Harpy-like systems for new tasks. Harpy's developers note that Harpy involved some very time consuming tasks, including making pronunciation dictionaries, manually tailoring the juncture rules, and compiling the pronunciation network. They note that some knowledge such as juncture rules, pronunciation variability, and duration effects are not easily represented in a graph and require considerable ingenuity. However, their ingenuity and successful

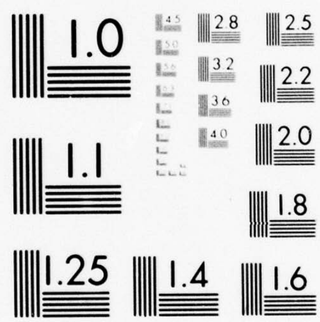
AD-A066 161 SPEECH COMMUNICATIONS RESEARCH LAB LOS ANGELES CA F/G 17/2
REVIEW OF THE ARPA SUR PROJECT AND SURVEY OF CURRENT TECHNOLOGY--ETC(U)
JAN 79 W A LEA, J E SHOUP N00014-77-C-0570

UNCLASSIFIED NL

2 OF 2

AD
A066161

END
DATE
FILMED
5-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

developments have already solved most of these problems, and they are working on automatic procedures for obtaining pronunciation dictionaries and juncture rules, plus ways of incrementally compiling minor changes in the network.

Harpy and Dragon, and to some extent HEARSAY II, achieved accuracy primarily by heavily constraining the task handled. (Earlier systems developed at CMU, SDC, Lincoln Laboratories , and SRI also highly constrained the task to achieve more accuracy.) In fact, Wolf and Woods (1979) argue that Harpy only dealt with an actual branching factor of 10, compared to HWIM's 196. They point out that Harpy's grammar allows mostly sentences that ease the recognition task by starting with stressed words, and the grammar allows very few sentence pairs that are similar in wording (and thus few that are readily confusable). They claim Harpy's language is non-habitable, so it could not be effectively used in most practical human-computer interactions. They assert that the performance of the 1976 HWIM system is not a true indicator of its potential, since (a) no vocabulary above 500 words was run on the system until the final weeks before demonstration; (b) the acoustic phonetic recognition component was incomplete; (c) many individual components were functioning far below potential and none had been tuned to its full potenial; and (d) the prosodic component was not even tested, much less tuned. Using and debugging the system components "had not been completed when the time arrived to run the final performance test" (Wolf and Woods, 1979, Sec. 14-5.3). It is interesting that when we surveyed expert opinions about whether the ARPA SUR systems would have met the original specification if given one more year, it was agreed that HWIM, more than any other system, might have made it. All this may be rationalization or speculation, and to date there still is no proof that HWIM could achieve high accuracy on the complex task they undertook (or even on a more restricted task like Harpy's). All the ARPA SUR systems except Harpy have been "dormant" since the completion of the project.

Like HWIM, the HEARSAY II system was intended as a general purpose architecture with several separable knowledge sources. Its developers argue that changes in individual components or system structure are relatively easy to accomplish, with the system and its various sources of knowledge readily evaluated. HEARSAY II was successful (90% accurate) on the document retrieval task that Harpy used, and is expected to be applicable to larger tasks such as HWIM attempted. Yet, it is difficult to comparatively evaluate HEARSAY II and HWIM. HEARSAY II was much closer to being fine tuned than HWIM when it was demonstrated, but not as well tested and adjusted as Harpy. Its success makes it a safer bet than HWIM, but it could profit considerably by several ideas and techniques that were developed for HWIM, including: the uniform scoring procedure; the lexical decoding network; the versatile ATN grammar; and the shortfall density strategy (or some other admissible strategy).

There is no indisputable evidence that Harpy's language is nonhabitable, or that either that language or some similarly small language with minimal complexity cannot be effectively used by talkers performing limited practical tasks. Chapanis and his colleagues (1977; cf. also Lea, 1979a) have found that humans do quite well in cooperative problem solving when restricted to vocabularies of a few hundred words, and can accomplish tasks almost as well if restricted to short (presumable simple) sentences. We still

await any experimental evidence as to the "habitability" of any very small language, or the degree of degradation in human performance resulting from various complex constraints on what can be said. For many immediate practical applications, a Harpy-like language may prove adequate. It also should be noted that while Harpy did use a vocabulary with a few similar words that could occur in similar structures, that careful selection of vocabulary and potential structural confusions would seem to be a reasonable and desirable feature to incorporate in a working system which must assure high accuracy.

We conclude that Harpy is the best choice for a recognition system that must accurately and quickly handle one small well-designed task. For systems that might be used for multiple small tasks or one or more large task, the choice is between HWIM and HEARSAY II. HWIM seems to be particularly suitable for cases where the developer wants relatively fixed, traditional components and wishes to explore alternative control strategies (especially efficient admissible strategies). HEARSAY II is suitable where a relatively fixed system structure (knowledge sources and blackboard) is desired, along with an excellent ability to evaluate individual knowledge sources that can be readily changed. HEARSAY II does have some flexibility in control stategy via its knowledge-source activation and scheduling procedures and its separation of policy from mechanism (Erman and Lesser, 1979). Since there are many control strategies yet to be explored, and also many components or knowledge sources that are still in definite need of improvement (or, in some cases, that have not been tested at all), both general approaches seem of current interest. However, most experts we have surveyed are more concerned now about improved or new components or knowledge sources (especially at the "front end" of the system) than alternative strategies, so the HEARSAY II system would seem particularly appropriate for further studies. If such an effort were attempted, the best of the appropriate HWIM ideas should be incorporated (cf. Klatt, 1977). Clearly, the ARPA SUR project contributed several excellent system structures and alternative approaches for efficiently integrating diverse sources of incomplete knowledge.

A-1.3 Performance Evaluation

The original study report which formed the blueprint for the ARPA SUR project emphasized correct understanding, so that the performance of a system would be measured by its capability to interpret an utterance and respond appropriately, as opposed to its ability to exactly recognize subunits of the utterance, such as phonemes, words, etc. Despite common impressions to the contrary, this ultimate importance of machine response was not a new idea originating with the study group, but was evident in various earlier writings (e.g., Fry and Denes, 1958; Peterson, 1961; Lea, 1969, 1970a,b) and is a natural extension of the communication theoretic notion that the ultimate purpose of a communication is to induce a desired response in the hearer (cf. Shannon and Weaver, 1949; Dean, 1953). However, the study group properly focused attention not on unimportant errors in phonetic strings or wording, but on ultimate interpretations and responses.

An important contribution of the ARPA SUR work was the experimental verification of the value of syntactic and semantic constraints. For example, Figure A-2 shows, for various tasks handled by the HEARSAY I system, how the

accuracy of recognition was substantially increased when syntactic constraints were added to the acoustic information, and then increased further by the addition of semantics (cf. Reddy, 1973). These tests were readily accomplished in HEARSAY I because of its ability to operate with or without knowledge sources like syntax and semantics.

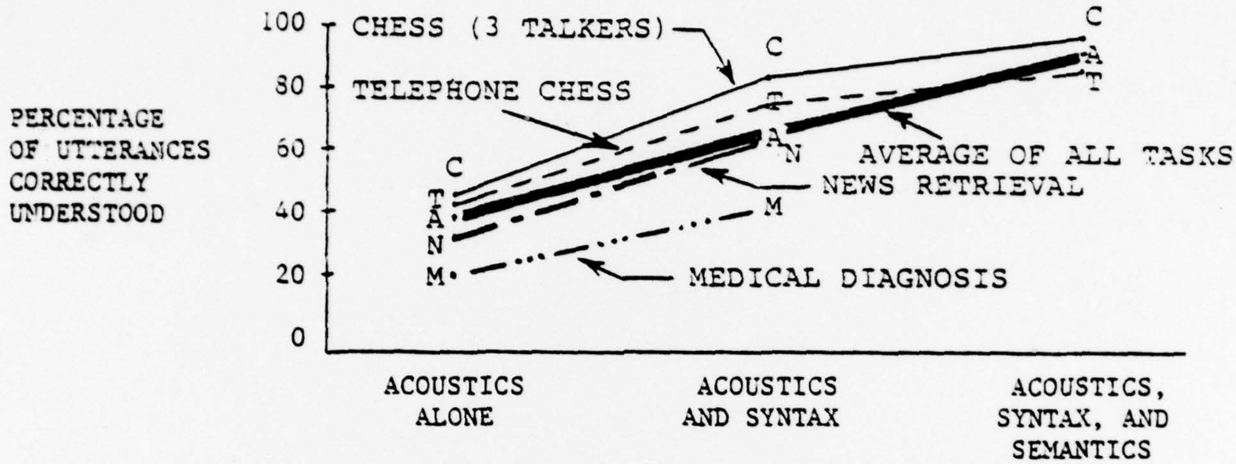
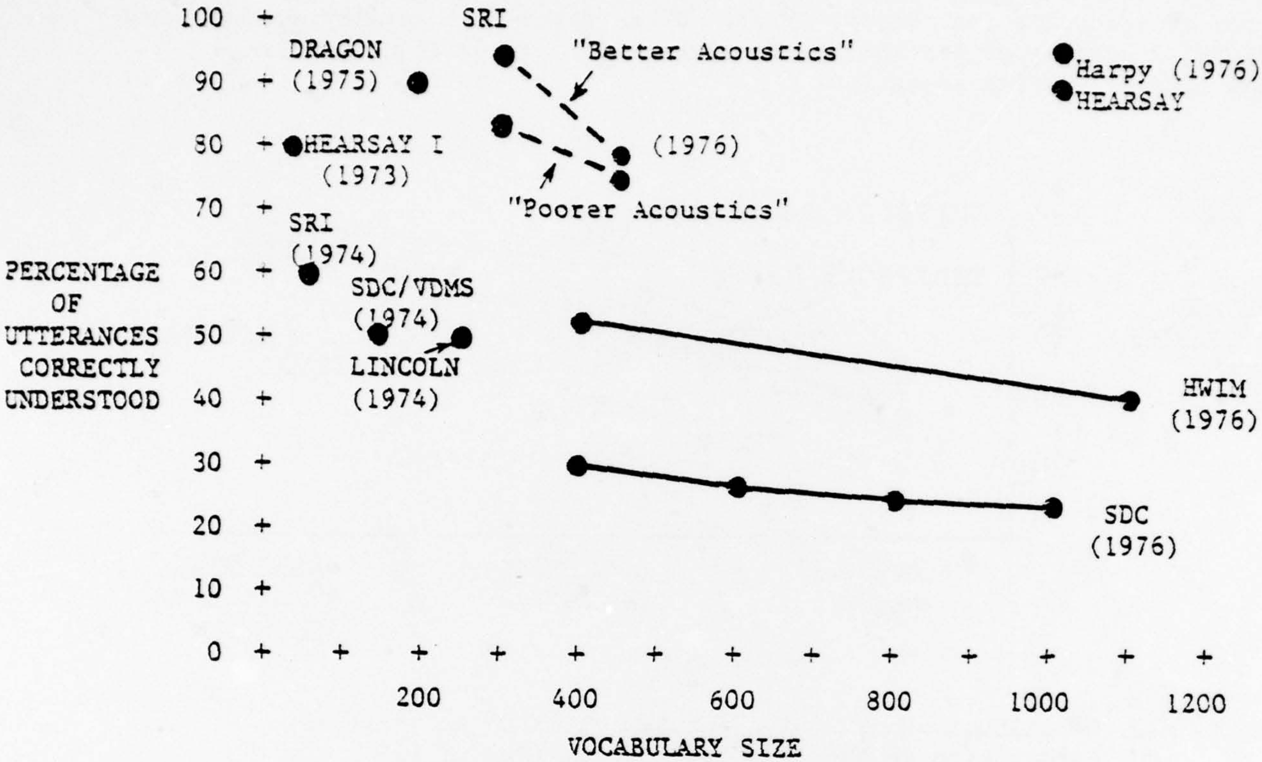


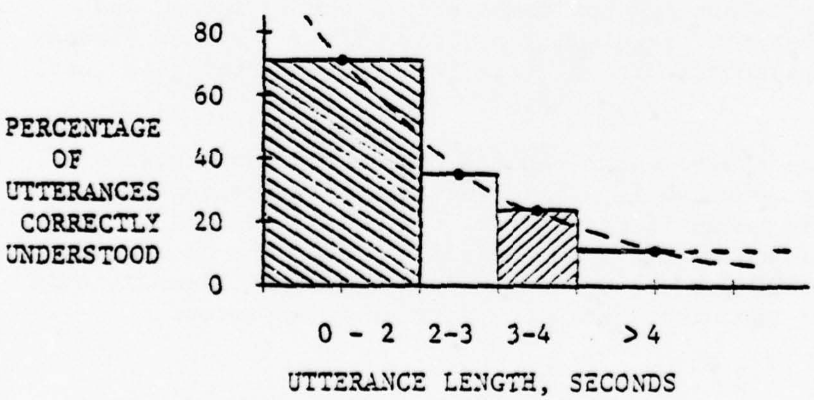
Figure A-2. Contributions of Syntax and Semantics to Accuracy of Recognition in the Hearsay I System (Reddy, 1973)

Such systematic tests of the contributions of various system components and task constrains are generally valuable. As Paxton noted (in Walker, et al., 1977, p. 973), "if it is worth building a system ..., it is certainly worth making an effort to understand how the system actually works, and experimentation is an important technique for doing this". We have already noted SRI's and BBN's experiments with alternative control strategies (Sec. A-1.2.2).

Figure A-3 shows some experimental results concerning how accuracy of understanding is affected by vocabulary size and utterance length, for various ARPA SUR systems. In general, for systems tested with varied vocabularies, the vocabulary size had only a slight influence on the accuracy of recognition, while, at least for one system, utterance length significantly affected the likelihood of the utterance being correctly understood,



(a) Effects of vocabulary size on recognition accuracy



(b) Effects of utterance length on recognition accuracy, for 1976 SDC system.

Figure A-3. Effects of vocabulary size and utterance length on speech understanding accuracy, for various ARPA SUR systems.

Perhaps more important is the evidence regarding how accuracy relates to the complexity of the language, as illustrated in Fig. A-4 by the error rate versus the static branching factor. Harpy and HEARSAY II seem to be highly sensitive to the static branching factor, while HWIM is not. An extension of the Harpy and HEARSAY II results to large branching factors like HWIM's 196 probably (but not with certainty) would have produced high error rates like those for HWIM. A valuable contribution of the ARPA SUR project was the development of the branching factor (as well as an old but largely unused measure of <u>entropy</u>; Goodman, 1976) as a measure of the complexity of a recognition task.

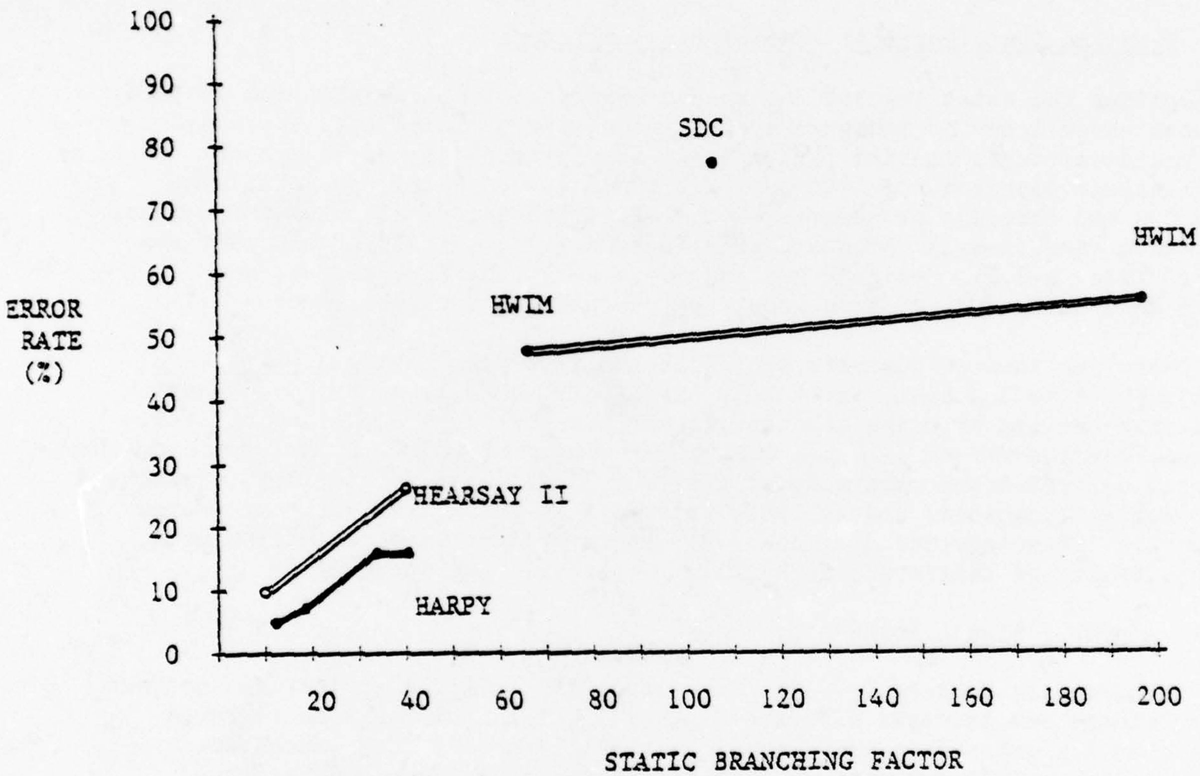


Figure A-4. Effects of static branching factor on recognition error rate

There are very few data points in these plots of recognition accuracy (or error rate) versus vocabulary, utterance length, and language complexity, and we would hope that further studies would provide more extensive evidence about what influences recognition accuracy.

In general, the utility of phonology, syntax, semantics, and pragmatic information is evident in the final recognition accuracy scores given for the ARPA SUR systems. While HARPY correctly labeled (by top choice) only about 42% of the allophones in sentences, it attained 97% correct <u>word</u> recognition by its network constraints on allowable pronunciations. Also, while .97 taken to the seventh power would have predicted a seven-word <u>sentence</u> recognition accuracy of only 81%, the use of linguistic constraints produced 91% having no errors in any of their recognized words. Finally, the distinction between having all the words right and having correct <u>understanding</u> of the utterance

was evident in that not just 91% but rather 95% of the sentences were correctly understood, so that the correct computer response would be possible, even if minor words were misidentified.

Similarly, the SCD mapper was able to recover from the fact that only 1 in 167 syntactically predicted words was correct by using a combination of phonetic, lexical, and syntactic constraints to reject 97% of all word predictions that were wrong. The long-sought-after goal of using linguistic constraints to aid acoustic information has finally become a reality.

A-2. Detailed Contributions: Components and Ideas

Besides the total systems for speech understanding, the ARPA SUR project produced several new or improved system components or "knowledge sources," and procedures or facilities for extensive analysis of connected speech. These include improved acoustic analysis tools (Sec. A-2.1), phonetic segmentation and labeling procedures (Sec. A-2.2), phonological rules and lexical processors (Sec. A-2.3), prosodic aids (Sec. A-2.4), syntactic analyzers and parsers (Sec. A-2.5), semantic and pragmatic analyses (Sec. A-2.6), and speech data bases and advanced knowledge of the speech signal (Sec. A-2.7).

Every new idea or analysis technique benefits from having a user that really puts it to the test, and the ARPA SUR work actually contributed many intangible results by using and testing earlier work from a variety of disciplines. Besides the work of the five groups involved in system building, special research contributions were made by the four "specialist contractors": Haskins Laboratories (phonetics and syllable studies); Speech Communication Research Laboratory (phonology and dictionaries); Sperry Univac (prosodic structures); and University of California at Berkeley (prosodics and phonology).

A-2.1 Acoustic Signal Processing

Although, as Neuburg (1975, p. 84) noted, the ARPA SUR project did not add any promising new acoustic parameters[1], still it did provide some improved techniques for extracting traditional ones. All the ARPA SUR contractors cooperated in a 1973 workshop on acoustic parameterization, and exchanged ideas and results of their various acoustic analyses of a common database of thirteen sentences which were pertinent to the various systems. From that workshop and later related interactions came decisions: to software preemphasize the speech to avoid erroneous low first formants in the LPC spectra; to use the Sperry Univac version of Sondhi's (1968) autocorrelation method of fundamental frequency tracking; to track formants by simple peak picking on LPC spectra, and incorporate smoothing and correcting procedures developed at Lincoln Laboratories (McCandless, 1974); to adopt BBN's "off-axis" LPC analysis and pole tracking to help resolve close formants; to detect voicing from band limited energies such as developed at Lincoln Laboratories; etc. SDC used and modified Haskin's "convex hull" procedure for locating syllabic nuclei (Mermelstein, 1975) and Sperry Univac's pitch tracker (Skinner, 1973; Gillman, 1975) and phrase boundary detection ideas (Lea, 1973; Bernstein, et al., 1976). Haskin's advocacy of the syllable as a useful recognition unit impacted the design of the SDC and HEARSAY II systems. CMU's earliest simple acoustic analyses with zero crossings were augmented by LPC and other parameters used by Lincoln, BBN, and other systems.

---

[1] Janet Baker's (1975) instantaneous frequency measure based on local zero crossing times might be considered a new acoustic parameter, but it apparently has had negligible impact on subsequent systems.

BBN (Makhoul and Cook, 1974) developed a selective method of LPC analysis, whereby fricative spectra were readily represented by a few poles, while sonorant sounds were analyzed with the usual many (12 to 15) poles. Klatt (1976) proposed a perceptually-based filter bank as a reasonable alternative to LPC analysis. CMU developed the simple ZAP DASH parameters and parameter-independent phonetic classifiers, and showed by comparative evaluations that several alternative sets of acoustic parameters gave nearly equivalent success in phonetic classification procedures (Goldberg, 1975). They also showed that PARCOR coefficients of LPC analysis were particularly attractive in speaker-independent recognition because, unlike most acoustic parameters, they could simply be additively averaged to obtain composite templates. Studies with semi-noisy computer rooms and telephone speech were acoustic advances that had rarely been attempted before.

General "dip detectors" (Weinstein, et al., 1975; Schwartz and Zue, 1976) were developed that can be used to detect potential boundaries between phonetic segments, at major changes in any acoustic parameter such as energy, formant frequency, spectral distance, etc.

A-2.2 Phonetic Segmentation and Labeling

Another basic parametric segmentation (preliminary to phonetic analysis) was involved in the detecting of syllabic nuclei from energy contours in systems at SDC, CMU, BBN, and Sperry Univac. Mermelstein (1975) of Haskins Laboratories provided a "convex hull" algorithm for locating about 93% of the syllabic nuclei, and his algorithm was incorporated into the SDC system. Sperry Univac's algorithm for locating syllabic nuclei (Lea, 1974; 1976) was tested on several speech data bases including a set of 255 sentences composed of only sonorant sounds (which represents a "worst-case" test of syllable detection capabilities), and the algorithm successfully located 91% of the syllabic nuclei, despite the lack of large energy dips in intervocalic sonorant consonants.

The syllable was a recognition unit in systems at CMU and SDC. The number of syllables in a word seems to be one reasonable metric of length, for normalizing word matching scores for words of different lengths. The number of syllables remaining in an utterance was also used by SDC and CMU as a length measure in rejecting hypotheses that could not properly account for the remaining data. Studies at SDC and SCRL suggested that syllable boundaries are too difficult to reliably locate to be used in recognition schemes, and in many phonological rules the syllable boundaries could be removed or ignored without altering the effects of the rule (Hanson, et al., 1976). We still await any hard evidence that phonetic constraints are less across syllable boundaries than within syllables, even though such evidence, if available, would help justify syllables as units in speech analysis.

New and advanced phonetic segmentation and labeling strategies were developed. Harpy incorporated the concept of "allophones" in its 98 sub-phonemic templates, and made a firm decision on the best (or closest) of these for each acoustic segment. Lowerre (1976) showed that Harpy was more efficient than Dragon in part because Harpy had phonetic segmentation. Harpy's pronunciation networks and 98 phonetic templates seem to have been successful enough to prompt Klatt (1977; also, Klatt, 1979b) to propose

that no strict phonetic segmentation of speech be done, and that sequences of spectral templates be used to represent words. The "folding" of word juncture rules directly into Harpy's pronunciation networks was also noteworthy. The full implications of Harpy's phonetic methods are not yet known, but they seem to have been instrumental in creating a trend towards allophonic template representations and de-emphasis of more traditional phonetic segmentation and labeling methods (cf. Klatt, 1977; Bahl, et al., 1978; Newell, 1978).

The high interest in Harpy's phonetic analysis procedures is not necessarily because of high accuracy in phonetic labeling. Harpy (and Hearsay II) only had the correct phonetic label as the top choice on 42% of its segments, or within the top three choices 65% of the time. (HWIM correctly identified the phonetic segment by the correct top choice 52% of the time, or within the top three choices 80% of the time, while SDC had the correct top choice about 50% of the time.) An interesting conclusion is that, with the proper linguistic constraints, Harpy was able to succeed in 95% correct sentence understanding despite only 42% correct phonetic labeling. One would hope that substantial improvements in phonetic labeling would still be possible; the IBM phonetic analysis obtains about 65% correct labeling (cf. Klatt, 1979b). Also, experiments at CMU (Shockey and Reddy, 1974) with phoneticians and linguists transcribing the phones in a language they did not understand suggest that about 60-70% of the phones should be machine transcribable, and more recent experiments showed reliable human transcriptions of 90% of the phones in people's names (wherein syntax, etc. does not help disambiguate, but the phonology of English is adhered to).

The phonetic lattice developed at BBN was an interesting contribution. It permits alternative conjectures as to what phonetic (or, quasi-phonemic) segments appear at various portions of the speech, and thus does not restrict word hypothesizing to only those words with the most likely ("nearest neighbor") phonetic segments in each region. However, that flexibility of alternative phonetic strings being preserved for higher levels can on occasion cause combinatoric explosions and confusions among alternative word sequences. Ambiguity arises not only from alternative segmentations (calling a region either one or two segments) but also from the assignment of a score for all possible segment categories at each position in the lattice. It has not been shown whether the ambiguity of the lattice helps or hurts phonetic recognition, and it is currently difficult to assess whether the lattice or Harpy's larger set of phonetic segments is better.

Despite the modest performance of the phonetic analyzers in the ARPA SUR systems, a number of definite improvements were made in detailed algorithms for detecting and identifying various speech sounds. Vowel identification procedures were improved, using speaker-dependent target frequencies for formants (at SDC, BBN, Sperry Univac, and Lincoln Laboratories). Formant trajectories were used for detecting, segmenting, and identifying glides, nasals and dipthongs (at Lincoln and BBN). Haskins Laboratories also developed a nasal detector (Mermelstein, 1975). New or improved algorithms were provided for detecting retroflexives, laterals, flapped dentals, intervocalic glottal stops, initial glottal fricatives, plosives, fricatives, and affricates, and for determining place of articulation. A common feature in all the systems was an initial segmentation into broad ("manner-of-articulation") classes followed by refined categorization.

The ARPA SUR project also produced a number of phonetic research tools that should be of interest in future work. The ARPABET developed during the project is a useful agreed-upon quasi-phonemic labeling scheme that is computer-compatible and yet phonemically reasonable. BBN developed an acoustic phonetic analysis facility that provides an interactive environment for rapidly performing a wide variety of acoustic phonetic and phonological experiments on a large data base of continuous speech. One can specify phonetic contexts and find all their occurrences in the data base, run algorithms like acoustic parameterization or segmentation and labeling procedures, perform computer-assisted hand labeling of continuous speech, display parametric results or scatter diagrams like $F_1 - F_2$ plots, and have the computer tabulate statistical results (Schwartz, 1976).

In general, while it is beyond the scope of this report to comparatively evaluate all the ARPA SUR techniques in acoustics, phonetics, lexical analysis, syntax, etc. it would be a valuable contribution to the field if someone provided such objective evaluations (cf. Part II of Lea, 1979a for some evaluations). An extensive effort was devoted, in particular, to phonetic identification procedures, and all that was learned should not have to be tediously relearned by new projects.

A-2.3 Phonological Rules and Lexical Analysis

One of the primary efforts of the ARPA SUR project was the compilation of over 200 phonological rules, and the selection of those that are most appropriate for use in systems. The rules selected were both analytic and generative in nature. They handled acoustic-phonetic, coarticulatory, junctural, and prosodic phenomena; but they did not handle derivational or morphophonemic information, since this knowledge was incorporated into the formation of the lexicons. The individual entries in the lexicons usually required more than one base form on which the rules would operate. The base forms were most useful if they were realizable, rather than abstract forms. The systems that were most successful were those that applied the rules before an entry was looked up. Indeed, in Harpy all the rules were precompiled into the network. The majority of rules were optional, rather than obligatory. Most of the rules included information regarding word boundaries, morpheme boundaries, and/or syllable boundaries, though studies were done to see which rules could be rewritten without boundary considerations.

The ARPA SUR project is the primary source of phonological rules used in speech recognition (along with the IBM system), so the reader is referred to (Shoup, 1979) for further details and example rules. A companion contribution was the development and refinement of phonological rule compilers and testers, at SDC, BBN, and SCRL.

A primary trend from the project was the combining of phonological variability with lexical representations. Harpy had a different path through its network for each alternative pronunciation of a word, and it also had junctural (word-boundary) phenomena incorporated as allowable transitions in the network. SDC had spelling graphs for words, which were similar to Harpy's full expansions of alternative pronunciations for words (though SDC included syllable boundaries as segments in its pronunciations). HEARSAY II used a Harpy-like pronunciation network in its WIZARD word verifier. A major

contribution in phonological and lexical analysis was HWIM's lexical decoding network, which merged common parts of pronunciations of different words and incorporated "wrap around" procedures for tying the ending of one word together with the beginning of the next word, allowing deletions and substitutions of elements based on coarticulatory regularities. Thus, as Klatt illustrates (1977; 1979a), optional deletions like "list some" sounding like "lissum" can be efficiently represented.

Lexical analysis procedures seem to have become quite accurate, though rather expensive in computer processing time. The SDC mapper in isolation missed only 7% of the correct words and gave 7% false alarms (with more misses on polysyllables and more false alarms on monosyllables, as might be expected). When operating within their system, it yielded 8% misses and 3% false alarms. Figure 12-10 and Table 12-5 in the chapter by Barnett, et al., (1979) about the SDC ARPA SUR effort suggests that while there is no combinatorial explosion of mapping time for longer words, the mapper consumes a large portion of the processing time of the SDC system. In the 1976 HEARSAY II system about 50 words of the 1011-vocabulary were generated at each syllabic nucleus position, and about 75% of the words actually spoken were correctly hypothesized. The word verifier in HEARSAY II accepted 94% of the correct words, but falsely accepted 49% of the incorrect words (cf. Klatt, 1979a, Table 11-4). The lexical retrieval process in HWIM found a correct word as the highest scoring word about 57% of the time and then used word verification to accept about 84% of the correct words, but also falsely accepted 34% of the incorrect words. Lexical retrieval in HWIM consumed 40% of its processing time, and word verification required another 15%.

Experiments showed that the system performances were not drastically degraded with increases in the size of the vocabulary (cf. Fig. A-3). The NOAH word hypothesizer introduced into HEARSAY II in 1977 performed with a degradation that was only logarithmic with vocabulary size in the full range of 500 to 19,000 words (cf. Chap. 7, Sec. 7-3.2.2.3 of Smith and Sambur, 1979).

Word verification came of age during the ARPA SUR project, and the parametric word verifier developed at BBN is considered one of the significant contributions (Klatt, 1977; also, 1979a). The verifier and synthesis procedures in HWIM apparently constitute the first attempt at actually using analysis-by-synthesis in a speech recognizer (Halle and Stevens, 1962).

Another significant BBN contribution related to lexical analysis is the uniform scoring procedure. Since all components of a system, especially the "front end", make hypotheses and offer priority "scores" about the content of an utterance, some means is needed for systematically combining the assessments of all components. This has usually been accomplished in an ad hoc manner, such as adding (or averaging) the scores on phonemes to get the score on a word, then adjusting scores based on prosodics, likely word sequences, etc. Each of the systems settled on a log likelihood method of scoring, and BBN carried that philosophy to a systematic combining of probabilities and the development of an admissible strategy for finding the best interpretation. Other system builders have since said they would like to incorporate a similar idea into their systems.

A-2.4 Prosodic Structures

Prosodic analysis facilities were among the new tools developed for speech recognition procedures by the ARPA SUR project. Besides the syllabification procedures mentioned previously, prosodic analysis algorithms included an improvement of Lea's earlier algorithm for detecting major phrase boundaries from fall-rise valleys in fundamental frequency contours (Lea, 1972; 1973a, b; 1976). The location of the F_o valley was shown to be just before the first stressed syllable of the following phrase. Another important algorithm that can be of use in various speech recognition systems is one that locates 89% of the perceived stressed syllables, from rising F_o contours, and long-duration high-energy syllabic nuclei (Lea, 1973a; Lea and Kloker, 1975). Less reliable but simpler algorithms for phrase boundary detection and stressed syllable location (and determination of rate of speech) were developed and tested at SDC (Bernstein, et al., 1976), but no data is available on their utility in the SDC (or any other) system. Harpy used minimum and maximum phonetic durations as conditions on the likelihood of a phonetic segment being present, but the significance of such prosodic information in improving performance was not studied.

In general, while the acoustic data and parameter extractions were available for determining important prosodic features within each of the systems, prosodic features played only a minimal role in the final systems. Despite the development of a procedure for using intonational phrase boundaries in the BBN parser, no system actually used prosodic features to determine large-unit linguistic structures and to aid syntactic parsing. However, the facilities are available for future use in studies of prosodic aids to speech understanding.

We reserve for Sec. A-2.7 discussions of experimental research conducted on prosodic regularities. For further discussion of prosodic aids to speech recognition, refer to Lea (1979d).

A-2.5 Syntax, Semantics, and Pragmatics

Also of interest were the distinct syntactic methods in the project, including the augmented transition network (ATN) grammar at BBN that combined syntax and semantics into "pragmatic" grammars, and the SRI "total language description" that focused on the grammar of speech as it is actually spoken, not as it is described in a textbook model. While the incorporating of semantic information into the category symbols of the pragmatic ATN grammar significantly improved the efficiency of the syntactic aspects of the HWIM system (Woods, et al., 1976, vol. IV), the resulting grammar was restricted to highly task-specific constructions. Thus, even though the ATN structure may handle versatile subsets of English (beyond finite-state or context-free grammars), the pragmatic grammars actually implemented are highly task-constrained. This is in line with the ARPA SUR goals of a constraining task. However, it then is a major task to develop a grammar that can handle a new task, unless it fortuitously uses constructions like the travel-management categories of "trips", "meetings", "fares", "sponsors", etc. Like Harpy, HWIM thus requires extensive effort to efficiently handle new tasks. However, new words and a few new constructions can be readily incorporated without recompiling the whole network, such as Harpy requires.

One of the major syntactic advances from the project was the ability to parse errorful strings, starting at arbitrary points in the utterance, so that parsing was not restricted to the standard left-to-right text processing methods. Word sequences could be parsed that did not form single non-terminal "constituents" in the grammar, and could be saved for use with later hypotheses. Hypothesized constructions ("islands") were then extended by predicting adjacent words either to the right or to the left, and appropriate words were filled in at short gaps between two non-contiguous portions of hypothesized structure (e.g., cf. Woods, et al., 1976; Reddy, et al., 1976). Not surprisingly, unconstrained right and left extensions of islands created combinatorial explosions of hypothesized word sequences, and strategies were (and are still) needed to constrain such island driving proceduring, such as only using highly reliable (especially multiple-word) islands, and controlling directions of extension (see Sec. A-12.2). Parsing in the presence of error requires abilities to handle and set priorities on alternative interpretations, and Harpy's beam search technique seems to be one of the best methods to come out of the ARPA SUR project. Harpy's pre-compiled integrated network was also particularly effective for constraining the possible word sequences to be tested. HEARSAY II used constraints on word pairs that can be time-adjacent to constrain sequences before higher-level syntactic analysis was done.

Semantic networks, discourse constraints, and task constraints were also included in the systems. For examples, HEARSAY I used chess board configurations plus legal and plausible moves, to help determine what was said in a voice chess game. Allowable "next steps" in the assembly of apparatus were to be used in the SRI/SDC system, and plausible things to say about trips, meetings, budgets, etc., were incorporated into HWIM's "TRIP" component. SRI conducted protocol analyses and used partitioned semantic networks and discourse information (such as antecedents for pronouns) in the development of their "top end" components for a speech understanding system. However, semantic and pragmatic constraints had little impact on the performance of the final ARPA SUR systems. Response generation was also included in the system designs, but had no significant impact on final performance and was not even demonstrated in most final system tests. It might be said that, besides handling the error and island driving involved with speech, the ARPA SUR efforts in higher-level linguistics primarily served as users of previous ideas rather than as creators of new concepts.

Of major importance for future work was the development of methods for measuring phonetic and lexical ambiguity and language and task complexity. The syntactic branching factor (Goodman, 1976), while not a totally adequate measure of language complexity, is a major improvement over earlier measures like vocabulary size and the number of syntactic productions or non-terminals in a grammar. Goodman's concept of measuring complexity by information-loss in a recognition "channel" is also intrinsic in later measures, like "entropy" (Sondhi and Levinson, 1977) and "perplexity" (Bahl, et al., 1978).

A-2.6 Experimental Research

Not all of the research results in the ARPA SUR project could be directly incorporated into computer programs during the rush to complete the final demonstration systems. Also, while it was not a primary objective of the project to extend the basic knowledge of the speech signal, several significant contributions came from experimental research conducted by the system-builders and the specialist contractors.

We have already noted in Sec. A-1 several experiments with system structures, contributions of various knowledge sources, and associations between complexity and recognition accuracy. BBN introduced an interesting concept of "incremental simulations" whereby humans simulate the functions of system components until their actions are understood sufficiently to be implemented (Woods, 1975). Simulation studies may continue to be useful in guiding future system designs.

BBN researchers also explored how humans use phonetic information in spectrograms to hypothesize words in sentences, when the spectrographic display is "windowed" within 300 ms segments to prevent syntactic and semantic context from aiding recognition (Klatt and Stevens, 1972). One third of all phonetic segments were transcribed without error, while another 40% were partially identified error-free. As might be expected, they found that stressed cardinal vowels, /i, a, u/, single prestressed voiceless consonants, and single nasals had particularly low error rates. This agrees with Sperry Univac's experiments showing that single sonorant consonants and phonetic segments in stressed syllables were more reliably categorized by five different automatic (machine) transcriptions than were unstressed or reduced vowels or obstruents (Lea, 1973c; Lea, 1979d, Fig. 8-1). Only 3% of the vowels were undetected in the spectrograms, which is confirmed by usual high accuracies in machine detection of syllabic nuclei and vowels. Front-back distinctions were difficult to reliably make, and this too has been evident in automatic vowel labeling programs, including in the final ARPA SUR systems.

The transcribers were then asked to identify the words in the spectrograms. Klatt and Stevens estimated that for words containing 4 or 5 phonemes, the probability of correctly identifying the word was only about 0.25 (or less). However, word identification was greatly assisted by having a computer offer a list of words whose sound structures were similar to the transcribed sequence, and the transcriber could then select the best word. (A similar procedure was later used in the SDC lexical subsetter.) Indeed, in every case where the computer hypothesized the correct word, it was recognized and accepted by the experimenter. These experiments supported the idea of introducing word verification into HWIM and other systems.

Haskins Laboratories also conducted experiments on human detection of words from spectrograms. Their method was to match portions of the spectrogram of a sentence with reference spectrograms for alternative words. They found that even when most words were incorrectly matched, the number of syllables in the hypothesized word sequence generally agreed with the actual sentence. Thus, syllabic units were reliably detected. Feedback of a reference spectrogram (that is, acoustic "verification") for each

hypothesized word did not help identify words unless most hypothesized words were already correct. As is now well-established, they found that manner and voicing of consonants were considerably more reliably detected than place of articulation. Haskin's researchers also explored perceptual guidelines for defining better distance measures for phonetic analysis (Mermelstein, 1976), and the dominant resonance in the front cavity of the talker's vocal tract was the best formant cue to place of articulation.

The ARPA SUR project required human transcriptions or judgments about the speech, as standards for evaluating machine results, so extensive work was done (primarily at SCRL and Sperry Univac) on repeatable and reliable techniques for providing orthographic, phonemic, phonetic, and prosodic transcriptions. The development of the ARPABET (cf. Shoup, 1979) as an agreed upon set of quasi-phonemic units was useful. SCRL transcribed 34 discourses orthographically and phonemically (using the ARPABET), and transcribed portions of 10 discourses in detailed phonetic form. SCRL also developed programs for generating subdictionaries, searching large databases or lexicons for various phonetic sequences or phonemes, or computing frequency of occurrence information within discourses. SCRL also selected final test sentences for the CMU and BBN systems.

A valuable part of the legacy from the project is the set of speech data bases compiled for various system tests and other purposes. CMU collected over 1,000 sentences, plus strings of digits. BBN and SDC each collected and processed several hundred sentences, and SDC and SRI collected 30 protocols of interactions. Other sentences were recorded and processed by intermediate systems like the Lincoln System, HEARSAY I, SPEECHLIS, and the SRI system. This listing is undoubtedly incomplete. Since designing and recording of speech data bases is one time consuming and expensive process in testing aspects of recognition, we would recommend to speech researchers that they consider using the extensive data bases compiled during ARPA SUR. Each system builder, of course, has to develop a data base to test his system, so that many sentences are usually recorded, transcribed, digitized, and processed. Some particular data bases that would be appropriate for "benchmark" tasks for testing future systems would be those used to test the successful HARPY system. To compare alternative systems, they need to be tested with equivalent tasks and the same speech data, since we currently do not know how to decide whether system A that yields 50% correct understanding on a difficult task is better or worse than system B that yields 90% correct understanding on an easy task.

Speech data bases were also developed to test specific components of systems, such as acoustic phonetic and prosodic tests made with 84 sentences (21 by each of four talkers) that phonetician Peter Ladefoged carefully transcribed (giving phonetic segments, stress levels, and time boundaries) for SDC. A large data base of 1100 sentences, each recorded by three talkers, was developed at Sperry Univac to study prosodic patterns and a few phonetic phenomena in various English sentence structures. SCRL compiled a large database of over 10,000 seconds of discourses, interviews, and separate sentences, and provided transcriptions for 30 protocols. It still remains an important task for the speech community to catalog all such speech data bases and make them available to other researchers.

One major area in which knowledge of the sound structure of spoken sentences was advanced was in prosodics. Besides developing computer programs for prosodic analysis (including algorithms for F_0 tracking, syllabification, intonational phrase boundary detection, and stressed syllable location), Sperry Univac (cf. Lea, 1976) also conducted experiments that demonstrated the following useful regularities:

- Automatic phonetic labeling schemes work more accurately in stressed syllables. Stressed syllables are islands of phonetic and phonemic reliability.
- Listeners can consistently determine which syllables in connected speech are stressed, and machine detection of stress could be expected to be (at best) 95% correct.
- Certain word categories (nouns, verbs, quantifiers, etc.) are consistently stressed, except that subordination (and coordination with repetition of parts of the structure) will decrease the perceived stress levels. Stress levels may help determine sentence structures.
- Cues to phrase boundaries are found in intonation "valleys" (fall-rise contours), unusually long intervals between stressed syllables, and phrase-final lengthening of vowels and sonorants. Distinct durations of pause, and large F_0 variations, occur at clause and sentence boundaries.
- Time intervals between stresses are a good measure of speech rate, and correlate well with frequency of occurrence of errors in phonetic labeling (i.e., indicate where phonological distortions occur).
- Intonational phrase boundaries could help select structural hypotheses in a syntactic parser, and showed promise of more quickly finding the correct parse in the HWIM system.
- A prosodically-guided speech understanding strategy was defined (Lea, Medress, and Skinner, 1975).

These prosodics studies substantially improved the evidence that prosodic information can be used to aid speech recognition, and defined explicit ways for using prosodics in aiding phonetic and phonological analysis, word selections, and parsing. Yet, we still await any substantial use of prosodic information in speech understanding systems, and extensive basic research about prosodic correlates of linguistic structure still must be undertaken (Lea, 1976: also, Lea, 1979d).

Further details about the technical contributions of the ARPA SUR project are to be found in the final reports of the various contractors (Bernstein, et al., 1976; Reddy, et al., 1976; Walker, et al., 1976; Woods, et al., 1976;

Cooper, 1976; Lea, 1976; Hanson, et al., 1976) and in publications and references listed therein. BBN produced (in the years 1974 to 1976 alone) 39 presentations and publications, six quarterly technical reports, and a five volume final report. Thirty five reports and publications were listed in the SDC final report. SRI listed 22 publications and reports, and CMU's list included 57 reports, 11 journal articles, 10 chapters in books, three conference proceedings and books, and 53 papers presented at conferences and workshops. The project also produced over 200 informal technical notes ("SUR notes").

APPENDIX B.

AN OPINION SURVEY

REGARDING

SPEECH UNDERSTANDING SYSTEMS

B-1. INTRODUCTION

We believed that in our efforts to define "gaps" in speech recognition technology and to define further work that is needed, we should consider the views of many other colleagues who have worked on various aspects of system design, evaluation, and application. Consequently, we prepared a lengthy questionnaire soliciting opinions about the impact of the ARPA SUR project, the best current techniques in recognition, the needs and gaps in the technology, the potential applications, and further work to be done. An announcement of the questionnaire was sent to about 160 colleagues who we knew had been involved in speech recognition work or related studies. The questionnaire was also announced at the December, 1977, meetings of the Acoustical Society of America (ASA), the American Association of Phonetic Sciences (AAPS), and the International Phonetic Sciences Congress (IPS-77), as well as at the 1977 Voice Technology Workshop jointly sponsored by the Naval Training Equipment Center, the Naval Air Development Center, and the National Aeronautics and Space Administration, Ames Research Center (Breaux, et al., 1978).

This appendix is a summary of the opinions expressed by the respondents to that questionnaire. At the end of the appendix is a copy of the actual questionnaire, with the composite results from all the respondents summarized for each multiple-choice question. The remainder of this appendix will be a summary of the implications of these survey results, a description of specific opinions and additional ideas written in by the respondents, and our evaluations of how all these responses compare with other reactions we have heard in our detailed discussions with speech recognition workers throughout the country. We also attempt to organize and encapsulate all this information into guidelines for future work.

We will structure our discussions as follows. In section B-2, we consider the background and qualifications of the respondents. Then, in section B-3, we summarize the general opinions about the ARPA SUR project. Next, in section B-4, we outline the opinions about the best current techniques in speech recognition. Current "gaps" in technology, and potential applications and market possibilities, are discussed in section B-5. Proposals for future work are considered in section B-6, and the overall evaluation of what this says about the future of speech recognition is given in section B-7. We believe, and the respondents overwhelmingly agreed, that the questionnaire was a good way to survey opinions about future speech recognition work that should be undertaken.

B-2. WHO RESPONDED?

An initial invitation to respond to the questionnaire was sent to 127 colleagues (46 in universities, 36 in government, 24 in industry, 13 in non-profit laboratories, and 8 abroad). Twenty six returned the form requesting to be involved in the opinion survey. The questionnaire was then sent to those interested colleagues, as well as to other colleagues who had informally expressed interest in responding or who were known to have either participated actively in guiding projects of the ARPA SUR program or who were currently involved in leading other speech recognition work. Later, the questionnaire was sent to all attendants at the Voice Technology Workshop (excluding duplicates) and to a number of interested scientists who had heard the announcements and expressed interest at the ASA, IPS-77, and AAPS meetings in Miami, in December, 1977. A total of 157 questionnaires were thus distributed, all to individuals who had some background and interests in speech recognition (19 in universities, 61 in government, 60 in industry, 12 in non-profit laboratories, and 5 abroad).

Frankly, the questionnaire was made long enough (32 pages) and detailed enough so that only those sincerely interested in evaluating previous studies and influencing further work in the field would respond. Similarly, individual questions were quite specific, requiring degrees of agreement or relative rank ordering, not only to get more detailed information, but also to encourage those who did not know too much about a topic to pass on to the text question. This was remarkably successful, in that there were frequent notes on specific pages saying "don't know" or "don't know enough to evaluate", or blank pages even though other pages were filled in quite completely. A spot check confirmed that unanswered questions were closely associated with areas that the respondents stated (on the first page of the questionnaire) were not among their areas of work experience or qualifications for future work.

A total of 34 responses were received, which was 22% of the number of questionnaires mailed, and which seems adequate to make the survey of general interest. We also have been told, by many of the 100 researchers with whom we have personally conferred, that they did not respond because they knew their opinions would be conveyed by their detailed personal discussions with us. Thus, the official 34 responses to the questionnaire are also extensively strengthened by many other judgments communicated in person. Those judgments are also reflected in the opinions and comments that we, the authors, present throughout this report.

As shown by the responses to question P1 (summarized on page Q1 of the questionnaire), the respondents were predominantly trained as electrical engineers or computer scientists, with considerably less background in speech sciences such as phonetics or linguistics. This is evident pictorially from the bar chart shown alongside question P1 on page Q1. (As shown with this question and others on page Q1, and throughout the Questionnaire, we have endeavored to explain alongside each question how we obtained the final composite results, and how they may be highlighted pictorially.)

The responses to question P2 (page Q1) showed that the respondents were quite an elite group of well-qualified opinion leaders, with an average of about 10 years experience in speech recognition or related disciplines. The solid line on page Q1 shows the average relative levels of work experience in various aspects of speech recognition. These results were compiled by having the respondents rank order the topics they had worked on most during their speech recognition work, then assigning a score of 10 points for each individual's highest-ranked (rank "1") area of previous work, 9 points for

second ranks, 8 points for thirds, etc., then averaging over all those who responded to this question (P3, on page Q1). As might be expected for a group dominated by engineers and computer scientists, primary areas of previous work were: first, control strategy and system design; then, segmentation and labelling; then, extraction of acoustic parameters; then, procedures for word matching and verification; etc. Least among the areas of expertise were general research, contract monitoring, prosodics, semantics, and scoring procedures (in ascending order).

The dashed line with question P-1 shows average scores obtained from the rank ordering for what the respondents feel most qualified to do for future work in speech recognition. This generally follows along with the solid graph of previous experience, but shows a general "modesty" of showing qualifications to be somewhat lower than experience. Prosodics, and segmentation and labelling, show somewhat prominent "gaps", in which qualifications for further work are considered to be quite a bit lower than previous experience. This might be said to be an admission that considerable new expertise needs to be brought to bear on those topics. While a lot of work has been done for decades on segmentation and labelling, the experts seem to be calling for new knowledge to adequately deal with future problems in both segmental analysis and the newer topic of prosodic analysis.

Sixteen (47%) of the respondents were participants in the ARPA SUR project, while eighteen (53%) were not (cf. question P4 on page Q2). No "party line" or bias on the ARPA SUR project should thus be dominating the responses to be reported here, but it only takes unanimity between ARPA SUR workers, and a few other agreeing respondents who did not participate in the ARPA SUR project, to yield votes favoring that project. The pooled responses to question P5, showing strong relative interest in speech understanding systems, do in part reflect the high percentage of ARPA SUR participants. Since, on some such questions, we can expect ARPA SUR participation to markedly influence opinions, we will distinguish between the responses of participants and non-participants. Thus, on question P5 (page Q2), we see that much more interest in speech understanding systems is evident for ARPA SUR participants than for non-participants. (We shall return to the question of the relative importance of various types of recognizers again, in section 5).

All but two of the respondents represented groups that were either currently engaged in a speech recognition project, involved in related research, or interested in working in speech recognition (cf. P6 on page Q2). Three fifths of the respondents are currently actively engaged in a speech understanding (or recognition) project, primarily working with specific aspects of speech understanding. A few are selling products or developing total systems. Approximately 40% are working as, or planning to work as, speech researchers, and a nearly equal number (39%) are primarily working as computer scientists. The next most common activities are as government contract monitors and government users of recognizers. Also included were commercial sources of recognizers, industrial users, and an applications consultant.

Analysis showed that only 37% of the respondents who had been ARPA SUR participants were currently involved in an active speech recognition project, whereas 78% of the non-ARPA workers were now active. Most of the ARPA SUR research groups have dispersed or lost key personnel, and many ARPA SUR leaders are now working in other fields such as speech compression and language analysis. This exit of well-trained workers from the field is one of the setbacks resulting from the abrupt end of the largest single source of funds in speech recognition. Responses show that those not now working on speech recognition would be interested in doing so.

In summary, we see that the respondents represent a good cross section of well-trained, long-experienced, and currently-active workers in speech recognition, who have participated in a variety of projects and specialities over a long span of years, and whose opinions are well worth our careful scrutiny.

B-3. OPINIONS ABOUT THE ARPA SUR PROJECT

B-3.1 What Was Its Value?

We will next consider how these experts assess the ARPA SUR Project. As shown by the responses to question 1 on page Q4, the majority of the expert respondents consider that the ARPA SUR project was of great importance, and no one thought it was not important. Comments specifically written on the questionnaires were that it "created a new field", showed the feasibility of speech understanding, and showed how much could be done with available knowledge. Another comment was that it was not important to speech science, as contrasted with computer applications (cf. Klatt, 1977; Lea and Shoup, 1977). As shown by question 2a (page Q4), the project was considered by 72% of the respondents to have been needed. Yet, only 50% considered it to be well-conceived. Specific criticisms were: that it was premature, since we are still knowledge-limited, not technology-limited; that strict performance goals were a bad diversion; that no intermediate demonstrable products were sought; and that it should have been recognized as a research problem, not a development task. A full 76% considered that the project goals and system specifications were (at least) ambitious. Two thirds considered that the ARPA SUR project produced a significant advancement bordering on a major breakthrough (27% considered it to actually be a major breakthrough; cf. question 10, page Q8).

B-3.2 Was ARPA SUR a Success?

Perhaps one of the most controversial questions that could be asked about any project would be whether it was a "success". Fifty three percent of the respondents (to question 3 on page Q5) considered the ARPA SUR project a success, and only 6 respondents (18%) disagreed to some degree. The opinions were mixed and noteworthy. One respondent said it demonstrated feasibility and showed where the complexities are. Another said it was a success, but did not go far enough. Others said it was "a success in learning but not in fully meeting specifications", a "success in producing a system but not in providing new scientific knowledge", a "limited success", "interesting work, but somethings didn't get done (e.g., acoustic phonetics)", etc. It was "not cost effective, but met its goals", according to another opinion. Others criticisms said that "in industry, the project director would have been fired", that the project "has delayed general acceptance of voice input technology", and that "knowledge was gained, but speech understanding has gotten a bad name in government". One respondent is "waiting for by-products and spin-offs" before final appraisal. "Success" had varied meanings.

Twelve respondents strongly agreed that the ARPA SUR project was a success, but ten of them were ARPA SUR participants. In fact, 12 (67%) of the 18 who agreed (to any degree) with its success were ARPA SUR participants. Another way to highlight this is to note that 75% of the ARPA SUR participants agreed with its success (none of them disagreed), while only 33% of non-participants agreed (33% disagreed).

The answers to question 4 suggest why the respondents did or did not consider the project a success. There is a direct correlation between

judgments of success and understanding about whether the ARPA SUR systems met or failed to meet the original design specifications. All 12 of those who strongly agreed that the project was a success believed the original specifications had been met (5 votes), slightly exceeded (5 votes), or well exceeded (2 votes). Four (67%) of the six who disagreed with the project being a success believed it had fallen somewhat short (1 vote) or well below (3 votes) the system specifications. Ten (29%) of the respondents were not sure that ARPA SUR was a success, and none of those respondents believed the specifications had been exceeded (only 4 thought they had even been met). Thus, "success" correlated with success in meeting specifications.

Comments on this issue were that meeting specifications was "a matter of interpretation", that "HARPY was good for satisfying goals, but it violated the spirit of speech understanding systems", that the "specifications neglected habitability", that the results were "very good for a long term research project", and the final results "still remain to be determined, but indications aren't great."

The respondents overwhelmingly considered that the meeting of specifications was important, with only four disagreeing. Two considered it the most significant result of ARPA SUR, but no one considered it the absolute criterion for evaluating the project. Twenty one (65%) considered it an important result of the ARPA SUR project. Anticipating the next question, several respondents said that the "technology developed was more important", that "the most important criterion was whether it <u>leads</u> to something or has a positive future impact", and that while meeting specifications was (for two respondents) the most significant result of ARPA SUR, it was the developed <u>capability</u> that was significant, not the demonstration of it. On related question 5 (page Q5), the majority (23, or 70%) considered the advancement in specific aspects of recognition were either much more important (45%) or somewhat more important (24%) than meeting specifications. One comment said "the techniques are the chief legacy, to improve future accuracy". Another stated that while some systems failed to meet specifications, they still each provided advances. Another noted that insight was gained on where the major problem areas are.

B-3.3 <u>What Were the Primary Areas of Contribution</u>?

The leftmost plot on page Q6 illustrates the respondents' views about the relative degree with which the ARPA SUR project <u>used previous knowledge</u> in various fields relevant to recognition. This plot was obtained by assigning 10 points to each respondent's highest-ranked field on question 6 (page Q6), 9 points to the next highest, and so on, then averaging over all respondents for each field. Thus the highest (rightmost) scores are those considered most effectively used, with <u>syntax and parsing</u> being most effectively used, followed in turn by <u>acoustic parameterization</u>, <u>segmentation and labelling</u>, <u>phonological rules</u>, <u>semantics</u>, and so forth. <u>Least</u> effectively used was knowledge about <u>applications</u>, <u>performance evaluation</u>, and <u>prosodics</u>. These conclusions are in agreement with the original goals of the project, since it was designed to apply higher-level linguistic knowledge such as syntax and semantics within the context of complex systems, following control strategies that permit effective interactions with more conventional recognition aspects of acoustic parameterization, segmentation, and labelling. The original project design mentioned no call for consideration of applications, but adequate performance evaluations could reasonably have been expected if the developers were to decide that the original design specifications had been met at final

demonstrations. However, the final systems were still being modified just before the final demonstrations in the fall of 1976, so that fully adequate performance evaluations were not accomplished before the project terminated. The system builders had hoped for an additional year of funding to make final system adjustments and performance evaluations. Another aspect of the low appraisal of usage of performance evaluation knowledge may be that the system builders did not make extensive use of component evaluations, ablation studies, operations research models, null and optimal models, causality analysis, or other performance evaluation techniques known to computer scientists (cf. Newell, 1975, pp. 36-45). Nor were the various systems tested on a common task or in other ways comparatively evaluated.

The other aspect of knowledge which did not get much use in the systems was prosodics. Of course, at the beginning of the project there was very little quantitative knowledge about prosodics that was readily translatable into recognition procedures (however, cf. Willems, 1972; Lea, 1972, 1973a,b). Prosodics was, and is, in need of extensive research before all its ramification could be applied to recognition. In addition, while research was done on prosodics during the project, very little of that knowledge was incorporated into the systems.

The rightmost plot on page Q-6 illustrates opinions about the relative significance of contributions to the various fields of recognition. It is our impression that this ranking (from question 7, page Q6) fairly well parallels the level of effort or amount of funding in these problem areas. Thus, work on control strategies and overall system structures was considered the area of the most significance contributions, and received extensive attention throughout the project. Next in order of significance is segmentation and labelling, followed in turn by word matching and verification, syntax and parsing, and phonological rules. At the bottom of this list are applications, performance evaluation, semantics, and scoring metrics, in an ascending order which (with the possible exception of semantics) is comparable to the smaller amounts of effort devoted to those topics.

One reasonable way to view the plots on page Q6 is that wherever the plot of contributions (the righthand plot) is significantly higher-valued than the previous information (the lefthand plot), there was a substantial contribution of new knowledge and capabilities in that field. Thus, new advances occurred in segmentation and labelling, phonological rules, prosodics, word matching and verification, scoring procedures and control strategies. In contrast, while much was used of previous knowledge in syntax and parsing (and in acoustic parameterization), the respondents seem to be acknowledging that no major new ideas or advances were made in those areas.

When given the opportunity (in question 8, page Q6) to make their own lists of the best technical contributions of the ARPA SUR project, the respondents most frequently (namely, in the lists of eight respondents) mentioned the concepts for organizing diverse sources of knowledge. Related contributions mentioned were control strategies (4 occurrences in the lists), system organization (3), and the HARPY organization and integrated network scheme (4). Phonological rules were mentioned 5 times, with word boundary effects in word matching (twice), and SDC within- and across-syllable rules for coarticulatory effects (once). Prosodic contributions were mentioned four times. Use and refinement of LPC analysis techniques was mentioned three times, with segmentation and labelling, acoustic phonetics, detection of syllables and nasals, and formant tracking each getting one mention. Word matching and verification received two votes, along with a mention of BBN's verification strategy, and a mention of BBN's tree-structured access to the lexicon. Also mentioned (once each) were parsing in uncertainty, direction-free parsing, higher level language processing, analytic syntax, and simply

parsing. Data bases of large size were noted also. Philosophical advances noted were that ARPA SUR created interest in the speech problem, dealt with connected speech, and demonstrated the feasibility of limited CSR (continuous speech recognition).

B-3.4 Were the Supportive Research Contracts Useful?

Four Support Contractors (Haskins Laboratories, Speech Communications Research Laboratory, Sperry Univac, and the University of California at Berkeley) conducted basic and applied research in speech sciences and were asked during the project to relate that work to the systems being developed. As shown in question 9(a) on page Q7, the respondents considered that work moderately important, with only two thinking it was of little importance. They (i.e., 82%) overwhelmingly agreed that the idea of independent research efforts was a good one. They had mixed opinions about whether the system builders should have done this research themselves, with more disagreeing than agreeing. They were generally uncertain about whether the topics studied had been among the most needful of study (58% uncertain, 32% agreeing, 10% disagreeing). They were approximately equally split about whether the support efforts should have been directed primarily towards: basic research relevant to speech understanding; development of new knowledge sources or recognition methods; or helping on problems defined by the system builders.

Comments regarding the supporting research efforts were that: they formed an essential ingredient of the project; they made the most important contributions, particularly in relation to level of funding; they added necessary manpower; they added breadth and experience; and they contributed valuable ideas and criticisms. Criticisms were that: they were not well integrated into the project; they should have been freer; they should have emphasized acoustic phonetics more; and they should have contributed directly to the systems from the very beginning. One respondent felt that there was "room for both activities" of (1) basic research with no immediate application and (2) direct help in solving specific problems defined by the system builders. Other respondents noted that: the lack of a follow-on project disallowed real use of the support work; the research should have been looking towards a follow-on effort; and five years is not enough for basic research fundamental to speech understanding systems.

One respondent made the following lengthy comments: "In the absence of an overall coordinated plan, the support efforts didn't seem to have a well defined place to fit in - in some cases it merely appeared that they were doing their own thing. This seemed to be more the fault of the project organization, mainly. This seemed to resolve itself, somewhat, in the latter years, as specific supporters worked out more or less close alliances with certain system builders. This led to some collaboration, which was good. But I'm not sure that such "favoritism" would be a good policy for another similar project - it would seem to limit the ways that supporters could affect the project."

B-3.5 How Good Were the Final Systems?

The respondents were asked (in question 11 on page Q8) to rate each of the four ARPA SUR final-demonstration systems, selecting from seven alternative descriptions. The plurality vote (36%) for the HARPY system was that it was a "good system, but with limitations". Other votes were: "excellent system, widespread utility" (18%); "good results, but wrong method" (15%); "excellent system, limited applicability" (12%), and "don't know enough to evaluate" (18%). HEARSAY II also had a plurality vote of "good system but with limitations" (45%), with other votes for "excellent system, limited

applicability" (10%); "excellent system, widespread applicability" (7%); and "don't know" (a big 38%). The plurality vote for the BBN HWIM system was "poor results, but good design" (40%, including three write in votes that are closest to this description); other votes for BBN were "don't know" (23%); "good system but with limitations" (13%); "excellent system, limited applicability" (10%), "excellent system, widespread utility" (7%), and one vote (3%) each for "good results but wrong method" and "wrong design". The SDC system had a plurality (majority) vote for "don't know enough to evaluate" (59%), with "wrong design" coming in second (22%), followed by "poor results, but good design" (11%) and single votes for "good system but with limitations" (4%) and "excellent system, widespread utility" (4%).

The general plurality conclusions are thus that CMU's Harpy and HEARSAY II are good systems with limitations, BBN's HWIM gave poor results but is a good design, and the SDC system can't be adequately evaluated (cf. Klatt, 1977).

Many specific comments were made about the systems, as briefly summarized here:
<u>Harpy</u>: It is the "key" to general SUS. It did work, so it is not all wrong. It gave good results, but is it a dead end? It has clear and significant limitations; it is reasonable for limited languages with well defined domains when training is possible; it is good only for constrained and regimented speech; its successful operation depends upon severe constraints on its grammar, hence it is not very habitable or adaptable; it has severe syntax limitation for query inputs; it apparently is not applicable to larger grammar applications; its diverging grammar effectively lowers the branching ratio, and it wouldn't work on higher word branching ratios. It is speaker-dependent. The Itakura distance metric is limited. No flexibility is left after building constraints into the network. It is primarily a <u>recognition</u> (not understanding) system; it doesn't factor knowledge properly. Dynamic programming is absolutely wrong when you want to understand as opposed to recognize. The system is limited by the high cost of expansions in vocabulary or grammar and by the user enrollment (i.e., training) required. It is too slow and needs to be more accurate. The acoustic phonetic processing is good, but the network is <u>bad</u> for big systems.

Notice that the comments about Harpy are overwhelmingly negative. Those that liked it had little more to say, but those that didn't like it had much to say, because it is the one getting widespread current attention, since it worked best.

<u>HEARSAY II</u>: It is a laboratory tool, more flexible (than HARPY). It contains good ideas, and is useful in future systems. It is primarily an understanding system. It is the only system that has the mechanisms to allow any and all knowledge sources needed in speech understanding. Its limited language constrains its applicability, but it has good potential. The blackboard idea is good, but needed more time for full development and testing. The knowledge sources are weak. It is too slow and needs improvements. The generality built into it was more than what was needed for its task. Its control strategy seems fuzzy and ad hoc.

In general, the comments about HEARSAY II were quite positive, with the specific criticisms seeming to be readily rectifiable.

<u>HWIM</u>: The basic design was right. It was a good laboratory tool with good and interesting features; it has the scope and flexibility to be profitably improved; it is intrinsically the most interesting and theoretically satisfying system; it has some strong components; its potential seems high. It had some good ideas, but unfortunately was not stabilized long enough to

make it work; it was too ambitious to be brought on line in time; it was not well tested. It is perhaps over-general (or over-ambitious); the syntax was too much and inhibited rather than helped; a branching factor of 200 is too hard. It is very slow, so that no comprehensive performance evaluation was done; it was not optimized and had too little constraint, thus being massive and slow. The shortfall scoring led to terrible control structures, so that some utterances took an hour to process. Some system integration was still needed; it was not really completed; many bugs needed to be fixed before a thorough performance evaluation. Its generality across speakers is not known. It had problems in segmentation and labelling; it should use a sequence of spectral models to define allophonic units. The project was haphazardly run.

In general, the respondents' reactions to the system design of HWIM were very positive, but the over-ambitious goals, lack of syntactic constraints, and delay in putting the system together were critical problems. The lesson of "freezing" a system early enough to permit extensive performance evaluation and adjustment is very apparent from this system development project.

SDC: The final SDC system had good acoustics; the bottom end was sensible and fairly powerful, but the basic top-down structure and overall system were too rigid and too weak. The bottom and top were not integrated and needed considerable work; the top was poorly conceptualized. It was limited, ad hoc, and top-down, so we can not learn much from it. The word sniffer was a good idea, but should have been done at the parametric level. The project problems (loss of a computer, and the break with SRI) put it at a severe disadvantage; it is hard to evaluate.

In summary, the respondents generally felt that SDC had done some good work in the acoustic phonetic ("bottom-end") aspects of recognition, but unfortunately the final system hastily put together after the loss of their computer and their inability to complete the system with an SRI-developed "top-end" caused the final result to be inconclusive and disappointing.

B-3.6 What Would One More Year Have Produced?

Since the low performance of at least two ARPA SUR systems was attributable to their not being completely integrated and tested before the final demonstrations, we asked the respondents (in question 12 on page Q9) whether the systems would have met the original ARPA SUR specifications if given one more year at similar funding levels. The majority vote for each system was that the respondents were uncertain (76% of those responding, for BBN; 87% for SDC; and 87% for an SRI-developed system). The majority (7, or 64%) of the eleven respondents about the BBN system that had any certainty at all about whether or not it would succeed said that it would. Indeed, the BBN system received the only "yes" votes, except a single "uncertain, but probably" for the SDC system.

Specific comments about the BBN system's potential for success were that: it had problems in computation time, but might have met the other specs; it would need a 3000-to-1 speed up, plus some accuracy enhancements. One respondent said the major hindrance was segmentation and labelling, and it was otherwise close to specifications, so it should be possible in one year. Another said he did not know if one year was enough. One was concerned about the BBN group's philosophy of always adding components rather than freezing the design.

The bottom end of the SDC system would have been in order if it hadn't been stopped short, and a usable top end might have been developed despite the miles between SRI and SDC. No other positive comments were made about the potential of either the SDC system or an SRI system. The SDC system was said to be a poor design, lacking flexibility and lacking a back-tracking ability. SRI had no front end, and would have speed and size as problems.

B-3.7 What about "Habitable" Languages for Computer Input?

One of the system design dimensions that was not fully defined in the original ARPA SUR specifications was the nature and utility of the language for interaction with the machine. Also, Harpy is criticized for being too restrictive in its language, while HWIM may have been too ambitious for the ARPA SUR project. A major issue then is: How much communicative ability must an interactive language have if it is to be really useful for foreseeable tasks?

The concept of "habitability" of languages is intended to judge relative utility of a language. A language is said to be "habitable" if users can readily learn it and keep within its constraints in actual conversations, so that spoken sentences are grammatical and otherwise correct and recognizable. We asked the respondents a series of questions (13a to d on page Q10) about the significance of habitability and how the ARPA SUR languages relate to a spectrum of ever-more-habitable languages.

A full 88% of the respondents considered the questions of language habitability for SUS's to be very important (41%) or important (47%), and only one (3%) considered it of little importance. The respondents primarily thought that habitability for foreseeable SUS's was either very difficult to attain in a moderate size effort (34%), difficult to attain (31%), or now within the scope of current capabilities (31%). The plurality vote was that habitable languages were somewhat higher than the best ARPA SUR capability (35%), although 27% considered them comparable to a moderate ARPA SUR capability. Other respondents noted that it is equal to the BBN language; or most like the early SDC system; or dependent upon the task and not to be confused with complexity of the language. Another comment said that real time operation was more important.

On the question of how habitable the ARPA SUR languages were, the BBN language received the highest number of votes as being habitable (6, or 29% of the responses to this question, compared to 2, or 10%, for Harpy and HEARSAY II, and none for SDC). On the average, its rating was "almost habitable". The plurality vote for CMU Harpy, HEARSAY II, and SDC was "far from habitable", with SDC getting the highest percentage of such votes of non-habitability (67%, compared to 62% for Harpy, 56% for HEARSAY II, and 36% for BBN). Comments were: that none is close to habitable though BBN is far beyond Harpy; that it is a myth to think a 1000-word system can be habitable; that it depends upon the task and how people use the system; and that Harpy is so restrictive one could ask about authors "from Copenhagen or Stockholm" but not "from Stockholm or Copenhagen".

B-4. OPINIONS ABOUT THE BEST CURRENT TECHNIQUES

Our survey extended beyond evaluating the ARPA SUR project, to a study of the total current technology in speech recognition. In this section, we summarize the opinions of the respondents about a number of general and specific topics appropriate to establishing the best current speech recognition techniques. Since the respondents have a total of around 350 years

of experience in this field, their opinions about the best current techniques deserve careful consideration.

B-4.1 Which Systems Are Most Relevant to Future Work?

When these experts were asked to rank the available (continuous) speech recognizers in terms of their relevance to the way future work should be done, their responses (to question 1 on page Q 11) showed Harpy the most relevant, followed in turn by HEARSAY II, and a near-tie between HWIM and the IBM system. ARPA SUR participants ranked HWIM above the IBM system, while non-participants interchanged that ranking. Intermediate ARPA SUR systems that were among the moderately-ranked systems included the Lincoln Laboratory System and SPEECHLIS. Other systems given some recognition were an SRI-developed system and the Univac system. Write-in choices included isolated word recognizers (particularly Threshold Technology's), the January 1976 SRI-SDC demonstration system, and LOCUST (a refinement of Harpy for small computers).

In explaining their choices, the respondents made many relevant comments about the various systems. Harpy was considered: way out in front on all counts; indicative of a technology and should be exploited in the next phase; an excellent choice as an initial base for further work and worthy of exploring to find out how extensible it is and how well it can be made to handle the tasks it is best at. The Locust extension of Harpy was considered by one respondent to be the current best bet for small real applications. IBM's system was considered by one respondent to be like Harpy in many ways, and another respondent noted that it worked much better than Harpy on Harpy's grammar. HEARSAY II's architecture was considered unique for handling multiple knowledge sources, but others considered it was not promising compared to Harpy and that the distributed-independent control of the Blackboard system was not a very good idea. HWIM was said to have the right total design, to be the most general and potentially most extensible to hard problems, to have good dictionary back up procedures and an interesting hierarchical control in its uniform scoring strategy, and to need further evaluation of its semantic knowledge. One considered the SRI-SDC framework and ideas good, but in need of new ideas, and, as another noted, good acoustics. Other systems like the Bell Laboratories systems were noted to be relevant at one level (presumably the level of restricted word sequence recognition) but not at other levels (such as for complex speech understanding tasks). One respondent suggested that future systems should solve increasingly-difficult speech understanding problems that clearly use linguistic and task constraints to correct phonetic errors and word-matching errors.

On a related question (number 5, on page Q 13), the respondents showed their consistency by favoring Harpy, HEARSAY II, HWIM, and the IBM system for good methods for the "next generation" of systems. The non-ARPA participants, however, ranked the prosodically-guided system (Lea, Medress, and Skinner, 1975) higher than the IBM system. In fact, the top-ranked system for the non-ARPA respondents was the prosodically-guided system.

B-4.2 What about Search Strategies and Island Driving?

When asked which search strategy is best in speech understanding systems, the respondents favored following the best few paths first, either with backtracking or without backtracking (such as Harpy worked). Noone favored breadth first analysis, but two favored the optimal test of all alternatives. Thus a beam-search strategy such as Harpy's would

be recommended, though some comments favored the guaranteed best answer through a HWIM-like "shortfall" strategy, or a best-first strategy. One noted that it really depends upon the quality of the acoustic segmentation and labeling, which when good would make the best path first with backtracking very good, but when bad it is better to use a bounded breadth strategy.

"Island driving and the use of islands of reliability" was considered by a large majority to be useful provided the islands are truly reliable. One comment suggested what is important is that the system be aware of the reliability of any portion of the input. One respondent observed that while island driving sounds good, the problem is that it doesn't constrain further choices sufficiently just to know island contents, and it is better to insure that you can get past some bad regions. Other comments suggest that island driving is a yet-to-be-proved hypothesis, and depends on the system architecture and search strategy. Island driving apparently needs a supplemental strategy to keep the alternative hypotheses down to a reasonable number.

B-4.3 What are Some of the Best Recognition Techniques?

Before considering general evaluations of which system components need further work (questions 4 and 5 on page Q 13, question 2 on page Q 19, and question 6 on page Q 22), we will continue our summary of evaluations of detailed aspects of current technology. One indicator of which techniques the respondents consider best is their own listings of the most significant contributions of the systems they personally worked on. ARPA SUR respondents mentioned (on question 6, page Q 14): Harpy's integrated representation of knowledge sources and search strategy; HWIM's uniform scoring philosophy (mentioned three times), lexical decoding network (mentioned three times), incremental simulation efforts, and parametric word verifier; and SRI's language definition facility, performance grammar, syntax for a wide range of "English in use", and semantic and discourse analysis. Also mentioned, but not clearly identified with a specific system, were: the general framework and representations for acoustics, syntax, semantics, etc.; top-down word-matching; new control strategies (four mentions); formulation of explicit phonological rules; methodology for experimenting with system control alternatives; acoustic-phonetic recognition techniques; and methodology for evaluating a system. Non-ARPA respondents mentioned dynamic programming and the viterbi algorithm, low-cost recognizers, accomodation of parallel conflicting hypotheses in the system, structuring of a comprehensive partitioned set of knowledge types, and an equipment-independent semantically and syntactically based system for real-time interactive use. Some mentioned the general efforts towards handling real-time continuous speech and improved understanding of tasks or applications as contributions.

B-4.3.1 What Techniques Would You Borrow? - When asked what ideas they would now incorporate from other systems that they had not worked on, the ARPA SUR respondents mentioned Harpy's integration of knowledge sources, HWIM's scoring philosophy (mentioned twice), HWIM's lexical decoding network and phonological rules used in lexical matching (mentioned twice); HWIM's parametric word verification; Harpy-like word verification; BBN's APR (restructured); Harpy-like phonetic-level labeling; HEARSAY II's multiword seed techniques for island driving; "some of HEARSAY II"; HEARSAY II parallel processing concepts; and semantic interpretation of partially completed theories. Non-ARPA respondents mentioned the linear programming technique of Dragon, Locus, Harpy; the Harpy processing rules; the word recognition technique of Texas Instruments; and the prosodic information

of Univac.

These ideas are in substantial agreement with our discussion in Sections 4 and 5 of this report regarding promising methods to use in future work.

B-4.3.2 How Should Segmentation and Labeling be Done? - The respondents are not in favor of using individual 10 ms time samples (such as IBM's latest system does), without a phonetic segmentation. They most strongly favor the lattice of phonetic alternatives (a la HWIM), but also approve somewhat of: a large set of alternative phonetic segments (a la Harpy), English phonemes, and diphones, in decreasing order. They are slightly opposed to syllables as unanalyzed units and strongly opposed to the idea of having no segmentation below the word level. Also mentioned were: arbitrary-sized units (data-driven); use of phrase effects; word boundaries; a hierarchy of segments (sub-phonemic to syllable size); and allophone units, each defined in terms of a time sequence of spectra and other parameters. In summary, they suggest that segmentation should be done, and the phonetic lattice is a preferred way to accomplish it.

B-4.3.3 What Prosodic Information Should be Used? - Respondents were in favor of extracting and using various types of prosodic information, with highest agreement about locations of stressed syllables, numbers of syllables, syntactic pauses, and detection of intonational phrase boundaries. One write-in vote appeared for detection of word boundaries.

B-4.3.4 What Lexical Techniques Should be Used? - The respondents considered that the basic dictionary or lexicon in a speech understanding system should not have only one entry per word, but should have baseforms stated in phonemic forms and have a priori probabilities assigned to alternative pronunciations. They were generally undecided about whether the dictionary should be fully expanded so no rules are necessary to specify expected pronunciations, whether a syllabary of expected pronunciations should be used, or whether or not to have word templates of acoustic data. Several comments suggested that this was dependent upon the system design, and one suggested that the English phrase (and not words) should be the principal lexical item.

B-4.3.5 What about Word Verification? - The respondents were in agreement (see page Q 16) that word verification was useful, though they were generally uncertain about whether it is currently effective, best done with phonetic comparisons or acoustic comparisons (somewhat favored), and best done only when the word is hypothesized with a high score. They were somewhat opposed to doing verification of every word at every possible point in an utterance. It appears word verifying is a useful but not yet fully settled aspect of speech understanding systems.

B-4.3.6 What about Syntactic Analysis? - There was little conviction (cf. page Q 17) reflected about alternative syntactic analysis techniques, with the most agreement about it being combined with semantics, in a pragmatic grammar, and about the value of using a "best-few-first" beam search strategy. Again, this depends upon the system design. One comment said syntax should not be used to drive understanding, but only give weight to constrictions.

B-4.4 What are the Most Important Advantages of Voice Input?

An important aspect of current technology is concerned with why speech is a good modality for communication with computers. The respondents considered that, among the fifteen listed advantages of voice input (question 4 on page Q 25), those that were most important include the facts that speech is: the human's most natural communication modality; possible while the talker is mobile; possible in conjunction with other modalities; easy to use without training; the human's highest-capacity output channel; possible without visual contact; compatible with telephones and other channels; and demonstrated to an integral part of the most effective multimodal communication links. Inadvertently left off this list in the questionnaire was the advantage of freeing hands and eyes for other activities, but the responses reflected the importance of such an advantage, and this advantage was explicitly noted in one comment. Also noted were that speech is the best modality during stress, and is the only channel in which humans communicate spontaneously and without planning.

Respondents did not agree that speech input is important for detecting a talker's physical and emotional state. Also of little significance was the idea of speech being unaffected by weightlessness or g-forces.

B-5. OPINIONS ABOUT GAPS IN SPEECH RECOGNITION TECHNOLOGY

We come now to a primary reason for the survey; namely, to obtain expert opinions about primary problem areas in speech recognition technology, which would presumably warrent further funded work. Some questions regarding "gaps" in the field were intentionally duplicated or stated in several different forms, to establish consistency of conclusions, and to seek more detailed (as well as global) views of what work remains to be done. We will consider here opinions about what types of systems are most needed now (Section B-5.1), what aspects or components of recognition systems are most in need of further work (Section B-5.2), what specific aspects of acoustic and phonetic analysis need most attention (Section B-5.3), what applications need the most attention (Section B-5.4), and how big the market for various recognizers would be expected to be (Section B-5.4). In essence, this helps define the priorities for further work and the "bottlenecks" that need early attention, and sets the stage for defining recommended programs to fill these important gaps in current technology.

B-5.1 What Types of Systems are Most Needed Now?

We saw from question P 5 (page Q 2) that the respondents themselves are most interested in recognizers of constrained word sequences and digit strings, versatile speech understanding systems, and (for non-ARPA respondents) large-vocabulary isolated word recognizers. We also saw from question 1 on page Q 11 (cf. section B-4.1) that Harpy, HEARSAY II, and the HWIM and IBM systems (and possible a prosodically-guided system; cf. question 5 on page Q 13) were considered particularly relevant to future work. We would expect these preferences to be reflected in the respondents' rankings of the types of speech recognition or understanding systems that are most needed now. Indeed, the average ranks (for question 1, page Q 18) favored recognizers of digit strings and constrained word sequences, low cost isolated word recognizers, and fairly-well-restricted sentence understanding systems. Also ranked high (particularly by non-ARPA respondents) were recognizers of syntactically-constrained sequences of isolated words.

In response to a related question (page Q 23) regarding the utility of systems for recognizing constrained sequences of isolated words, the consensus seemed to be that while such isolated words may be better for the machine than connected speech, it would be a fairly difficult mode of interaction for the human to use. Some indicated that it might be in need of further study. As expected, non-ARPA respondents called for more immediate, limited recognizers, while top preferences for ARPA participants were recognizers of connected word sequences and restricted speech understanding systems (like Harpy).

In summary, these opinions suggest early interest in recognizers of several levels of capability, including, in decreasing order of preference:

- Connected word sequences and digit strings;
- Isolated word recognizers with syntax and low cost; and
- Restricted speech understanding (Harpy-like) systems.

We have advocated in our recommendations (cf. Section 5 of this report) the addition of research speech understanding systems to this list of systems worthy of early attention, to provide an ever-expanding capability in the field.

One respondent commented that we need a successful, practical system to demonstrate the potential for the long range unrestricted systems, and another considered low cost, high performance (low error) systems to be most important, and vocabulary size to be least important. Several said that each of the types of recognizers is needed in some appropriate applications.

B-5.2 What are the Primary Gaps in Current Technology?

We previously skipped over one of the early questions in the questionnaire, which asked which knowledge sources or SUS components should be given particular attention for improvement (question 4, on page Q 13). Here we will consider that question, and other related questions which provide consistent evidence about the respondents convictions concerning the most important gaps in current speech recognition technology. As shown on page Q 13, the respondents considered the following to be among the primary topics or components to deserve attention (in decreasing order):

- Handling coarticulation effects;
- Segmentation and labeling methods;
- Word boundary effects and rules;
- Phonological rules; and
- Prosodics

Notice that these are all aspects of the "front end" of a recognition system. It is interesting that after five years or more of work on higher level linguistic analysis, the focus for further advances should again be back on the acoustic, phonetic, phonological, and prosodic areas.

Specifically mentioned "gaps" also included fast speech rules, speaker normalization procedures, control and system organization, scoring of competing acoustic hypotheses, and a generalization of the syntax — semantics interface.

On a similar question (number 2 on page Q 19), a more detailed list of aspects of recognition was given, and the respondents were asked to indicate

which were the biggest (or most troublesome) "bottlenecks" or gaps in recognition. Two alternative scoring or ranking procedures were permitted: high, medium, or low; or a relative ranking (1=most important, 2=next, etc., with 26 alternatives offered). Scoring the responses as shown on page Q 19, and averaging, we found that the top one-third of their list of highest-ranked "bottlenects" were (in descending order, again):

- Acoustic phonetic rules;
- Segmentation procedures;
- Labeling of segments;
- Phonological rules;
- Acoustic parameterization techniques;
- Amount of data processed through the system during design;
- Prosodic aids to parsing;
- Sizes of segments; and
- Phonetic or spectral distance measures.

Again, it is evident that the "front end" of the recognizer is given priority, with a very comparable set of problem areas.

In specific comments, respondents to this question on page Q 19 also listed other gaps in word boundary detection, sentence boundary detection, sources of variability (ideolectal, dialectal, etc.) and speaker independence, telephone input, and system response time and practicality. Some suggested adequacies in current capabilities. One commented that acoustic parameter work seems pretty much done, while another thought segmentation was well in hand, and a third noted that dictionary entries are important but known. Language statistics were also said to be known and of unknown utility, and two comments suggested that use of syntactic constraints is reasonably well in hand and done quite well. Parsing was also considered "fairly good", and selections of task domains and user models were considered "not the issue". Amount of data processed was assessed as "crucial", but "solutions exist". In addition, there were frequent comments expanding on (or making specific) the items in the list. For examples: Itakura's distance metric was said to have inadequacies; improvements in segmentation and labeling were mentioned several times as crucial to other aspects of the system; many phonological rules were said to not be known; automatic learning of dictionary entries, and generalizations across speakers were considered necessary; fast, exhaustive word analysis is needed; efficient semantic analysis that is separable from pragmatic grammars was called for; performance metrics were considered crucial; and research on acoustic phonetic characteristics of English sentences was considered "definitely our bottleneck".

Finally, a third question (number 6 on page 22) permitted a general assessment of speech recognition needs, without such a detailed list, but with more opportunity to show the degree of respondents' convictions. These were specifically called "the most significant "gaps" (problem areas in need of further work)", and included the following top-priority half of the list (in descending order of strength of agreement):

- Acoustic phonetic analysis (the "front end" of a system);
- System tuning or adjustment based on extensive data processed;
- Measures and methods for performance evaluation;
- Fast or near-real-time operation;
- Prosodic cues to linguistic structures;
- Effective use of higher-level linguistic information to

constrain ambiguities; and
- Word verification.

Again, acoustic phonetic analysis and prosodics are among the top-ranked areas (indeed, for non-ARPA respondents, they are the top two ranked of the "gaps"). Surprisingly low on this list is the topic of phonological rules. This may be due in part to the inclusion of performance-oriented goals in this list, which were not included in any such degree in the other questions. One respondent specifically commented on the need for better "lower level" or "front end" aspects of systems, and considered that higher levels will be tailored to overcome phonetic errors. Another respondent argued that until acoustic phonetic analysis is much improved, recognition accuracy will not improve significantly. Other respondents noted the need for (a) speaker normalization and (b) statistical modelling.

We may conclude that acoustic phonetic, phonological, and prosodic aspects of systems need to be given early and prominent attention, and that systems and projects should be designed to assure fast testing with extensive data, using advanced methods of performance evaluation.

B-5.3 What are the Gaps in System Components?

Other questions were provided to define more detailed adequacies and gaps in the acoustic phonetic processing aspects of recognition. In question 3 on page Q 20, the respondents were asked to rank order various sets of acoustic parameters in terms of their utility in speech understanding. Top-ranked acoustic analysis tools included (in descending order: the first three formant frequencies; fundamental frequency; linear predictive analysis, and the amplitude or energy measure. Poles of the linear predictive spectrum were also moderately rated, but the total frequency spectrum, time domain analyses, zero crossing counts, formant amplitudes and bandwidths, and instantaneous frequency were not considered very useful.

B-5.3.1 Which Acoustic Parameter Techniques are Adequate or Need Further Work? - When asked (page Q 20) which acoustic parameter extraction processes need further work, the respondents agreed that primary attention should be given to the following (in descending order of priority): distance measures; vowel formant transitions and contours; and selection of acoustic parameters. Formant tracking, voicing decisions, fundamental frequency tracking, and vowel formant target frequencies were also somewhat agreed to as in need of further work, but use of LPC poles was not.

The conclusion would seem to be that analysis of contours of LPC-smoothed formant frequencies, fundamental frequency, and energy should be continued with priority attention, and other parameters and refined algorithms should be investigated.

B-5.3.2 Which Segmentation and Labeling Procedures Need Further Work? - Question 5 on page Q 21 inquired about the respondents' agreement with the relative need for work on various segmentation and labeling procedures. The phonetic analysis needs that were most agreed upon were (in descending degree of agreed-upon importance):

- Distinguishing among stop consonants (p,t,k,b,d,g);
- Distinguishing among nasals (m,n,ŋ);
- Detecting non-sibilant (weak) fricatives (f, θ, v, ð, h);
- Detecting laterals (l);

- Distinguishing among sibilants (s, ʃ, z, ʒ);
- Detecting nasals;
- Detecting and identifying glides (w, y); and
- Locating syllabic nuclei.

This is only the top one-third or so of the list, but <u>all</u> detections and identifications of phonetic units were agreed to in varying degrees. Those least agreed to are either fairly well accomplished currently (such as detecting vowels, retroflexives, or stops) or else of uncertain value (such as diphones). In general, much work is needed in this area.

Not reflected in the alternative answers to question 5 on page Q 21 is the possibility of <u>not</u> segmenting and labeling in the usual manner. We saw from question 8 on page Q 15 that segmentation <u>should</u> be done, and the phonetic lattice is a preferred way of accomplishing it. However, sub-phonemic units like Harpy used, which are essentially spectral patterns to be matched with a large inventory of alternative acoustic phonetic <u>templates</u>, are an alternative to the more traditional phonetic segments (cf. Klatt, 1977; 1979).

B-5.4 What Applications Need Most Attention?

We saw from the discussion in Section B-5.1, that the respondents favor the development of recognizers of digit strings and constrained word sequences, low-cost isolated word recognizers, and fairly-well-restricted speech understanding systems. Non-ARPA respondents were particularly supportive of the restricted recognizers that could be applied in the near future. In question 2 on page Q 23, the respondents were given the opportunity to indicate which <u>applications</u> for speech recognizers they agreed were very important and appropriate for future work. Their strongest agreements were for the following (in descending order): data retrieval; air traffic control; inventorying; command and control systems; package sorting; and cartography. Each of these is an area in which recognizers have been previously applied (or proposed), with some effectiveness. Least agreed-upon were gun fire control and spotting of key words in context, perhaps due to the military implications of those tasks or the difficulty of analyzing the expected noisy speech. The differences in preferences of the ARPA and non-ARPA participants reflect the different long-range versus short-range, and research-oriented versus applications-oriented, viewpoints of the two groups (cf. page Q 23). Write-in comments also mentioned applications in automated telephone transactions and speech aids for the handicapped.

B-5.5 How Big is the Potential Market?

The experts polled are not market analysts, and hence are not necessarily well qualified to accurately assess the market in speech recognition. Yet, they are knowledgeable about various available systems, government and commercial applications, and computer trends, so that their view on the market might be of some interest. In question 3 on page Q 24, they were asked how large a market they would estimate there to be (now or in the next few years) for various types of recognizers, based on their knowledge of the current state of technology and the various applications. These were given choices on an exponential scale (1, 10, 100, 1000, 10000, etc., units per year) and so in tabulating and averaging the results, such combinations as four votes for 100 units per year and two votes for 1000 units per year had to be summarized

by adding (and averaging) logarithms, then taking antilogarithms to obtain geometric means. The mathmatical procedures are shown on page Q 24.

The net results were that 5 or 6 thousand digit string recognizers, about 3 thousand small isolated word recognizers, over one thousand connected-word-sequence recognizers, and hundreds of each other type of recognizer might be expected to be marketed each year, for a total of over 12 thousand recognizers per year. Since the large bulk of these would cost perhaps several thousand dollars, and the more sophisticated speech understanding systems would presumably be much more expensive, this would project to a market in the order of $50 million per year. This is a substantial increase from the 200 or so that were sold, for a total of $6 million, in the first five years that commercial recognizers were on the market. We know of no signs that the market is opening up in quite the degree suggested by the poll, although over 500 hobbyist devices were sold in their first year on the market. The respondents' predictions would seem to be very bold, but conservative in comparison to the ten-year $1.5 billion market projected by one marketing consultant (Nye, Chapter 20 in Lea, 1979).

Respondents noted that the number of systems to be sold is a function of the costs, capabilities, and "acceptability" to users.

B-6. RECOMMENDATIONS FOR FURTHER WORK

It might be taken as a foregone conclusion that the respondents would recommend filling those primary gaps in current technology that were defined by their responses summarized in Section B-5. Here we consider specific opinions about further work that should be done, beginning (in Section B-6.1) with their views about the organizational features and system design specifications that would be appropriate if another large-scale project were undertaken. We consider (in Section B-6.2) their recommendations regarding general emphases and funding levels for further work, and their general remarks about future work (Section B-6.3).

B-6.1 What about Another Large Scale Project?

Two complex questions were asked of the respondents (on pages Q 26 and Q 27) concerning the best features for another large-scale project in speech understanding system development, if such a project were to be undertaken. The features to be evaluated were based on features of the ARPA SUR project, and also based on issues previously expressed to us about the best way to undertake further work.

B-6.1.1 What Organizational Features? - The expert respondents strongly favored (1) an extensive performance evaluation study as a part of the speech understanding system development, and also (2) having close interactions and interchanges among contractors, if another large scale project were undertaken (cf. page Q 26). Other organizational features favored were (in descending order of agreement): mid-term preliminary evaluations of systems; support contractors to do related research; and construction of several alternative systems. They were less certain (but still slightly favorable towards): strict deadlines for system demonstration; a steering committee for program management; and completely defined system specifications. ARPA SUR participants were slightly opposed to elimination of poorest systems at mid-term of the project, but non-ARPA participants were somewhat more in favor of such mid-term selections. The only listed feature that was opposed was close attention to immediate military applications.

Perhaps the reason why both groups opposed attention to immediate military needs is in the word "immediate", since their responses throughout the questionaire suggest that immediate needs are more identifiable with quite restricted word sequence recognition, whereas speech understanding is recognized as a longer-range research topic.

3-6.1.2 What System Specifications? - On a related question (page Q 27) of what system design choices they would include if another large scale project were undertaken, the respondents predominantly favored the following system specifications:

- Vocabularies of several hundred words;
- Speaker populations of about 10 to 100 speakers;
- Several forms of speech (in descending order of choice):
  - Sentences related to a restricted task;
  - Almost any sentence in a versatile subset of English;
  - Strictly formatted word sequences (and digit strings);
- Either normal computer rooms, or high-quality or noisy channels, with almost any transducer (close-talking microphone, telephone, good quality microphone, or radio channel);
  (This might be summarized as: practical operational input channels.)
- Slight tuning to the speaker, using a small set of utterences;
- A fair amount of use of syntactic constraints;
- Substantial use of semantic, pragmatic, and user model constraints;
- Very desirable to have real-time operation;
- A moderate (PDP-10-size) computer;
- Accuracy from 95 to over 99%;
- Around a 4 year time-scale, or unrestricted duration;
- With tasks such as command and control, information retrieval, or voice programming.

The overall differences from the ARPA SUR project are: less rigid demands on schedule and competitive developments; somewhat more practical environmental conditions; more ability (and testing) with various speakers; a spectrum of tasks of various complexities and forms of speech; higher accuracy; and more thorough evaluations.

Detailed comments on the various system choices suggested that: further work should involve both "small" and "large" systems; the vocabulary should be expandable by the user; the system(s) should be gracefully adaptive (and dynamically adjustable; 3 mentions) to new speakers; a subset of English syntax defined by the user should be used; continuous speech consisting of more than one sentence should be considered; the telephone should be used (five mentions); "real world" conditions where the user is located should be used; syntax and semantics should be used as much as is "appropriate"; real-time operation is very desirable for small "practical" systems but not important for large "ambitious" ones; the system should be made fast enough to permit large amounts of testing (10, or less than 25, times real time); a combination of a large computer (or multiprocessor) and special purpose hardware (high-speed signal processor, or LISP machines) should be used (three mentions); accuracy should be determined by the task; the project length should either be "open-ended" with careful project monitoring, or of indefinite length until the problem is solved, or 5 to 10 years, or a low effort over a longer time period, or "100 years" in duration; and tasks undertaken could be library retrieval work, data retrieval, voice numerical control of machines, fairly broad tasks with English speech, voice programming with FORTRAN, or involve a "bare bones" system expandable by the user.

Other comments suggested emphasis of the performance of subsystems, and the encouraging of modularity for ready exchange of successful subsystems. One respondent observed that the ARPA SUR project specifications "forced system builders to bite off more than they could chew", and another project "should allow for a user expandable system" and specify one or more well defined tasks for testing all systems. Another respondent called for "a 3 part program: (A) Fairly simple system, involving design choices and development but no components dependent on further research; (B) an ARPA like high performance system, not fully designable at the start but not unrealistic either; (C) Research on underlying problems and promising possibilities for incorporation in a B-type system if they work out". Another respondent also advocated both a low performance system and one for which developers "pull out the stops, but don't have too strong expectations about how far progress will be made". He considered that definite system specifications were "not so important on this round".

B-6.2 What General Emphases and Funding Levels?

In response to question 3 on page 29, the respondents recommended that work on speech recognition or understanding systems be either "increased somewhat in funding and effort" or "stimulated by a large increase in funding and effort". One respondent commented that the current funding level in the field is way too small. Noone recommended that work be ended now or sharply curtailed. A moderate increase seems to be called for. On the question of how the work should be funded, the vote was almost equally split between (a) supporting many separate modest projects, for various applications, agencies, and distinct purposes, or (b) supporting many modest projects that are carefully coordinated (the top choice, by a slight margin); or (c) combining current modest projects with a large coordinated project like ARPA SUR. Specific suggestions given were: that there be several projects of various sizes, some with long term and others with shorter term goals; that one or more substantial prime contractors, with a broad array of modest sized and diverse efforts, be funded; that coordination and interactions should be fostered, such as through a joint government committee and regular contacts at professional meetings; and that long-range, goal-directed basic studies are needed to solve the most difficult problems.

The vote (on question 3 (c), page Q 29) was about equally split between considering the future work to be: (first choice) applied research and development (in which recognition techniques are being developed and applied to develop prototype systems); or applied research (in which major questions of recognition techniques still need to be answered before applications are addressed); or long-range research (on speech characteristics, for later applications); or a combination of these. The best answer would seem to be that some of each is needed, as is reflected in our recommendations in Section 5 of this report.

We thus have the (no doubt expected) conclusion that considerable further work is needed, ranging from research to applications.

B-6.3 General Remarks about Future Work

The questionnaire ended with some general "essay" questions about the ARPA SUR project, current technology, and future needs, and these questions may be useful for the reader to ponder further. Here we summarize some of the final comments provided by the respondents.

One respondent noted that it is not really known how well the ARPA systems performed, and statistical evaluations deserve more attention. Another believed "a great deal more could have been accomplished", and would like to know why more specific results and scientific formulations of problems were not obtained. He gave his own answers, by criticizing the lack of intermediate demonstration systems and specific intermediate goals, and questioning the management by committee. Other detailed criticisms by the respondent are as follows:

"This is not an original comment, but ARPA did a great deal of harm to the development and demonstration of limited capability voice systems, because funding was frozen till ARPA "finished." Much was learned technically from ARPA/SUR, but it set back applications about 5 years, and was probably the most poorly coordinated major effort ever conducted by a DOD agency with respect to accessability of results to potential users (not other researchers). A lot of people had fun for a while but technology didn't move very far in terms of off-the-shelf capability. As a manager (predominantly) I am appalled at the total disregard for basic marketing that was deployed. The world of defense R & D won't wait 10 years to see if something will work.

"I think individual developers did pretty much what they were told and expected to do. The work was, within these boundaries, first-rate in most cases. The principal fault was ARPA management; as with virtually all ARPA efforts, the products/program cannot be assessed or used by anyone else including other DOD activaters.

"The most critical need for speech/voice system now is to demonstrate the capability to do real, useful things, practical tasks, before asking for additional money for basic research to improve capability. For virtually all military applications, computer core and speech is restricted. Hardware is needed to reduce the gigantic computer requirement to ROM, chips, solid-state ruggedized devices, etc. This should be done before improving systems, otherwise there will never be enough defense R & D dollars to achieve the needed improvements in recognition/understanding technology.

"It's too bad that the original and continued emphasis was provided by computer science and artificial intelligence. This led to an arrogant disregard of some of the most crucial problems, primarily acoustic phonetics, which they thought could be bypassed by sufficient upper-end intelligence. "Speech" people continuously tried to make the case for the fundamental speech-related issues, but with a classic "not-invented-until-invented-here" attitude, one could hear simultaneously that bypassing phonetics was a "breakthrough" and that the need for it was a "discovery."

"Ten years ago the bottleneck in ASR and SU was in acoustic analysis, and, through the great effort placed on it, acoustic analysis is now one of the best-developed ASR-related areas. Although analysis needs to be improved, the next bottleneck is to interpret the parameters once we've got them. This is the problem of acoustic phonetic analysis. There is every reason to believe that it, too, could, in a decade, be as well developed as acoustic analysis is today. And, just as signal processing is now a tool for studying acoustic phonetics, in a decade we may see phonetic analyzers serving as tools for the study of phonological rules. (I see prosodies as being a part of acoustic phonetics, broadly conceived). As the bottlenecks are resolved, the problems at other levels should become better defined and more

accessible.

"One final point on the same issue: as "outsiders", the speech and phonetics people were sometimes treated as "experts" in the sense that the system builders would have been delighted to have improved phonetics, if only the "experts" would spell out for them all the information and algorithms needed. Thus it may have been a disciplinary bias that would account for the view of phonetics as a trivial or unimportant area that could be handled by using at most "expert" consultation, rather than as a major area requiring extensive fundamental study."

Another respondent had this to say about the ARPA SUR contributions:

"I have been distressed by the emphasis given by many to meeting the specs' as opposed to the specific contributions the Speech Understanding Program has made to Artificial Intelligence, Linguistics, and Acoustic-phonetic research. This is probably due at least in part to the fact that it is hard to measure "success" short of a system we can all walk up and talk to, but I'll try to mention a few of the contributions I feel are significant.

"Contributions to Artificial Intelligence include development of new strategies and techniques for organizing large systems, for dealing with diverse kinds of information, for dealing with large "search spaces" often filled with uncertain information, and for handling multiple hypotheses about possible interpretations of data. (These are not mutually exclusive). The work at all the research sites has made contributions in this area. Another contribution to AI has been in the area of representation of knowledge, for example the use and refinement of semantic networks at SRI and BBN.

"In linguistics, there have been new ways of representing "grammars", new techniques for parsing (SRI & BBN), new ways of combining syntax and semantics, for example the SRI language definition facility and the BBN "pragmatic grammar". (These are, of course, two very different ways of combining them, with different goals.) There has been investigation of the role of "prosodics" and other information in syntax, although it has not been incorporated into any running system. And there has been the development of a representation for dialog structure and development of techniques for using it in interpreting inputs. Discourse is an area of growing interest in Linguistics, and several contributions were made by SU systems.

"All this is in addition to the acoustic contributions, which I am in a poorer position to evaluate, but am certain there are some.

"One other notable contribution, which I fear will be lost (if it hasn't been already) is the "training" of scientists who have at least some expertise in (or knowledge of) all of the disciplines which were to be combined in the SU effort. One of the original intents of the program was to bring together experts and technologies from diverse fields who were to combine their resources in approaching the problem. It really took close to 5 years for these people to begin to understand each other. Unless they can be used effectively (and soon) in helping with the next round of SUSS, the next "generation" of researchers will have to start over again learning what each other is doing."

Another respondent listed these ARPA SUR contributions: "System design concepts that coordinate and integrate knowledge in large amounts from radically different kinds of sources; communication and understanding among people from different disciplines". This same respondent recommended maintaining the stress on system building, but suggested that future projects should provide for experimenting with alternative components and control strategies. He suggested that we "need to pursue multiple alternative approaches both at different levels of complexity and at each level—particularly important with more complex systems", and "need close communication among research groups to ensure consideration of alternatives". He also said "we really missed the boat by not spending some effort at real comparison and evaluation of systems; for instance BBN should have implemented the HARPY grammar. This would give us several points in one dimension, rather than several points in many dimensions."

Regarding future system developments, one respondent suggested configuring systems after the way humans communicate with each other. Another respondent suggested that "a radically different 'front end' phoneme recognizer is needed". Another said that, "It wasn't until ARPA put mega-bucks into automatic speech recognition that significant advances were made. What is needed is continued basic research. It's not cheap". Detailed final comments from one respondent suggested that

> "The future should be considered both in "short term" and in "long term". For the short term, many of the ideas and techniques can probably be engineered to "working" systems, mainly for unconnected speech which will be suitable for some applications. There should be continuing work on these, both to show that it can be done, and to gain experience in the installation and use of speech for input.
>
> "For the longer-term, there needs to continue to be a mixture of AI, Linguistic and Acoustic researchers working together on the problems, and learning from each other. Not only do individual problems need solving, but we need MUCH MORE experience in combining components together into systems and testing these systems to better understand how the various components interact (or don't) and what they contribute to the overall problem. One of the weaknesses in the first 5 years was that too much time was spent on the pieces, but I suspect that was due at least in part to the fact that the pieces were all harder to build than had been anticipated.
>
> "Techniques to be used in future systems can only be considered with respect to the goals and requirements of the systems. Some of the techniques developed for systems such as HARPY can be used in short-term systems with limited capabilities, but I do not think they would be advisable for much more general, long-term systems. The reverse is also true, some of the more sophisiticated language techniques, particularly used by BBN and SRI, will really only be useful for longer-term more elaborate systems.
>
> "I think that before any new systems are built, a careful study evaluating and comparing the various systems and their approaches should be made. This should be done by several people who span the range of knowledge included in the systems (AI, Acoustics, Linguistics) and they should extensively test and measure what is happening in the systems. Then, based on this background, new systems can be designed which, hopefully, reflect not only the experience of having built these earlier systems but also a more careful analysis of what went into them.
>
> "Any new round of speech research should continue to stress system building", but more strongly emphasize actually keeping the systems together and working out solutions to problems within the

context of the systems. There should be testing and evaluation throughout the work. This does not mean research should not go on in each of the "components", but it should be done in the context of the whole system."

We have quoted all these detailed comments here not necessarily because we agree with them or believe they were all fairly stated, but rather because the respondents devoted considerable time to have these opinions expressed, and their conclusions deserve consideration as much as any responses to multiple-choice questions.

B-7. SUMMARY OF THE SURVEY

In response to final questions (page Q 30), the respondents were somewhat in agreement with the idea that the survey was a good way to review the ARPA SUR project. They were uncertain about its value in surveying current work. However, they were in stronger general agreement about the value of the questionnaire, in surveying opinions about future work. Some difficulty in, and disagreement with, summarizing positions through multiple choice questions was noted, and, as expected, some questions were considered difficult to answer outside the context of a particular system. Some of the questions were considered by one respondent to be structured with a linguist's bias. One respondent remarked that the survey was useful in helping him clarify his own opinions on various aspects of the ARPA SUR project.

The survey was certainly helpful to us, and it guided our recommendations for the field, by extending beyond our own necessarily limited experience to that of ARPA SUR participants and non-participants working on various aspects of recognition. Lists and evaluations of ARPA SUR contributions were provided, the adequacies and inadequacies in current technology were clarified and their priorities determined, and promising techniques for further work were defined. Now the task for the field is to use this knowledge to effectively advance the capabilities and usefulness of speech recognition.

PART A: PROFESSIONAL INTERESTS

P1. I was educated as (check one best answer, or rank: 1 Primary, 2 Secondary, etc.)

A	N					A	N
7.0	4.5	11.5	Electrical Engineer	2.4	Mathematician	0.5	2.1
1.0	1.6	2.6	Linguist	0	Speech Clinician	0	0
0.7	2.5	3.2	Phonetician	___	Other: Psychologist 4.0;	0	4.0
8.0	4.0	12.0	Computer Scientist		Physicist 3.6; Other (Miscellaneous) 3.0	1.5;1	2.1;2

P2. Years actively involved in a discipline related to speech understanding systems (SUS)

A/N: 2 (0/2) 0-2, 6 (2/4) 2-5, 14 (9/5) 5-10, 8 (1/7) 10-20, 3 (3/0) over 20.

P3. Give a rank order of (a) what you worked on the most during your SUS-related work; and (b) what you consider yourself most qualified to do relevant to speech understanding or related projects (1 = most, 2 = less, 3 = even less, etc.; give at least three):

(a) Work Experience				(b) Qualifications for Future Work		
A_{15}	N_{18}	T	ASPECT OF WORK	A_{15}	N_{18}	T
4.73	3.56	4.09	Acoustic Parameterization	3.15	3.16	3.15
4.73	4.67	4.70	Segmentation and Labelling	.3.62	3.25	3.44
3.60	2.60	2.82	Phonological Rules	3.54	1.42	2.58
0.60	2.67	2.15	Prosodics	0	0.34	0.33
4.40	3.67	4.00	Word Matching and Verification	4.15	3.42	3.80
4.27	3.22	3.70	Syntax	3.00	2.00	2.52
2.53	2.39	2.45	Semantics	2.31	1.92	2.12
3.13	1.94	2.48	Scoring Procedures	3.00	1.50	2.28
6.40	3.83	5.0	Control Strategy, System Design	6.38	3.75	5.12
3.73	2.94	3.30	Performance Evaluation	3.08	4.17	3.60
0.87	6.00	3.67	Applications	3.15	6.67	4.84
0	2.50	1.36	Contract Monitoring (in Gov't. etc.)	0	3.42	1.64
		___	Research on Acoustic Phonetics; Speech Physiology;			
		___	Other: Voice Authentication; Speaker Differences;			
		___	Other: Digital Commun. Systems; Articulation;			
		___	Other: Man Machine Dialogue; Project Management; Articulatory-to-Acoustic Relationships.			

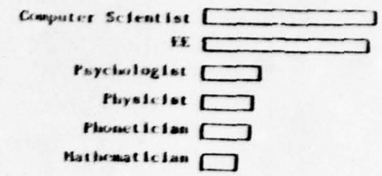
112

TABULATION PROCEDURES AND RESULTS

GENERAL NOTE: Throughout the questionnaire, A = ARPA SUR participants, N = Non-ARPA respondents

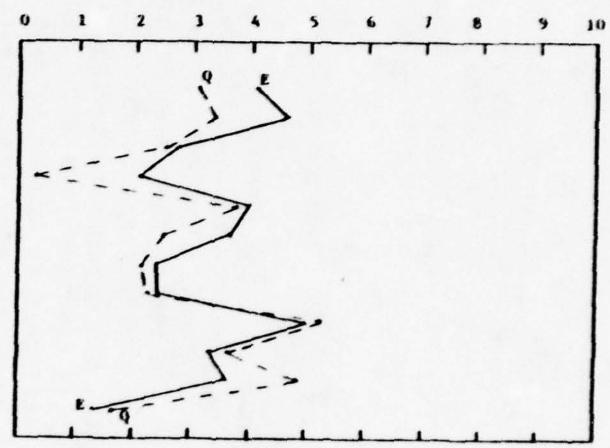
Total or averages for all respondents are listed under T.

P1: One point was assigned for each check mark or rank of 1(Primary); 0.5 point for rank 2; 0.2 point for rank 3; 0.1 for rank 4.



P2: 0-2 years was assigned as 1 year, 2-5 assigned 3 years, 5-10 assigned 8 years; 10-20 assigned 15 years; and over 20 assigned 23 years. The net result was an average of about 10 years (10.8 for ARPA, 8.8 for non-ARPA respondents).

P3: Top rank of 1 was assigned a score of 10 points; 2nd rank = 9 points; 3rd = 8; etc. The average score was then calculated, and the results for the total group of respondents are plotted below, with top-ranking topics on the right. E = work experience; Q = Qualifications for future work.



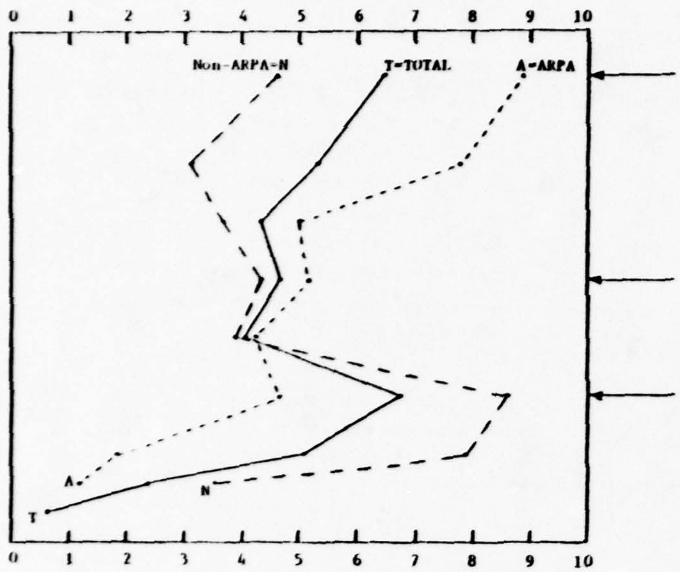
Primary areas of past work thus include: control strategy and system design; segmentation and labeling; acoustic parameterization; word matching and verification.

Primary qualifications for future work: control strategy and system design; applications; word matching and verification; performance evaluation; and segmentation and labeling.

Prominent "gaps" between previous experience and future qualifications are evident for prosodics and segmentation and labeling.

P4: 47% were ARPA SUR participants, and 53% were not.

P5: 1st rank = 10 points, 2nd = 9 points, 3rd = 8, etc. Averages were calculated and plotted for each group of participants.



P6: Some respondents checked more than one answer to this question. Most (78%) non-ARPA respondents are currently active in a SUS project, but only 6 (37%) of the 16 ARPA participants are now active in SUS projects.

P4. I <u>16</u> was / <u>18</u> was not a participant in the ARPA/SUR Project.

P5. Rank order your interest in the following types of speech recognition/understanding systems (1 = highest, 2 = lower, etc.)

A	N	T	
8.88	4.556	6.49	Versatile systems with habitable languages and full (or expandable) speech understanding capability;
7.81	3.11	5.32	Speech understanding systems of the highest capability sought during ARPA/SUR (i.e., 1000-word vocabulary, sizeable useful subset of English, ambitious task or tasks, multiple speakers, etc.)
5.000	3.67	4.29	Speech understanding systems of moderate ARPA/SUR-like capability
5.188	4.28	4.71	Speech understanding systems of the lower ARPA/SUR-like capabilities (continuous speech, restricted syntax, restricted task, etc.)
4.125	3.94	4.03	General connected-speech recognizers without use of syntax, semantics, or task constraints (a la IBM's system, etc.)
4.688	8.56	6.74	Connected word-sequence recognizers (formatted phrases or sentences, digit strings, etc.)
1.875	7.94	5.09	Large vocabulary (~300-word) isolated word recognizers
1.188	3.5	2.41	Small commercial devices for isolated word recognition
		0.656	Other: Narrow Band Speech Transmission Also: phonetic transcriber; Phoneme or allophone recognition

P6. My organization currently is:

A	N		
6	14	20	Actively engaged in a SUS project. We are
	3		a. 3 Selling Products
1	1		b. 2 Developing a SUS
4	8		c. 14 Working on aspects on SU
1	4		d. 5 Other: Applications; Installing voice data entry; Man/Machine Dialog
2	8	10	Involved in SUS-related speech research on the following topic(s) Acoustic Phonetics, Word Spotting, Syntax & Semantics, NB Sp XMTN; Training applications; Speaker differences & segmentation strategies; Applications
7	2	9	Not currently active in SUS, but interested in
4	2		a. 6 Developing a system
4	1		b. 5 Working on aspects of system development
4	1		c. 4 Planning a project in SU
1	0		d. 2 Other: General Language Understanding, Higher Level Processing
2	0	2	Not directly interested in SUS work

P7. I am currently working as, or planning to work as, a:

A	N		
	4	4.0	Government Contract Monitor
	2	2.0	Government User of ASR/SUR devices
7	7	14.0	Speech Researcher
9	3	12.0	Computer Scientist (primarily outside ASR/SUS)
1	2	1.3	Developer of SU devices(s)
	2	2.0	Commercial source of SU devices
1	3	4.0	Other: Language ID, Industrial user; Applic. consultant; commun. system developer

[BLANK SPACE REMOVED]

Q3

PART B: GENERAL OPINIONS ABOUT THE 5-YEAR $15 MILLION ARPA/SUR PROJECT

1. The ARPA/SUR project was:

A	N		
10	6	16	Of great importance to speech understanding technology
5	4	9	Of considerable importance
1	7	8	Somewhat important
			Not important
			Other: Created a new field; Showed how much could be done with available knowledge.

Comments: Major impact in visibility of SUS concepts; other goals might have contributed more;

2a. The ARPA/SUR project was at the outset

9	3	12	Well conceived and definitely needed
3	1	4	Well conceived but not needed
2	9	11	Much needed but ill-conceived
0	1	1	Ill-conceived and not needed
2	2	4	Other:

Comments: Too much attention on performance goals; goals were a bad diversion; intermediate-complexity demonstratable products were needed; should have been research, not development; needed relation to applications

2b. The original ARPA/SUR project goals and system specs were:

	6	6	Unreasonably ambitious
9	2	11	Very ambitious
2	7	9	Ambitious
5	1	6	About right
0	1	1	Somewhat modest
0	1	1	Too limited
			Far too limited

Q4

3. All things considered, the ARPA/SUR project was a success.

	Strongly Agree	Agree	Am Not Sure	Disagree	Strongly Disagree
A/N T:	10/2 12	2/4 6	4/6 10	0/5 5	0/1 1

Comments: "Meeting specs does not a successful project make"

4. As I understand it, the best ARPA/SUR system(s):

A	N		
2	0	2	Well exceeded
4	4	8	Slightly exceeded
7	4	11	Met
3	5	8	Fell somewhat short of
0	5	5	Fell well below
			Other:

the original design specifications. This is:

			The absolute criterion for evaluating the ARPA/SUR project
2		2	The most significant result of ARPA/SUR
12	7	19	An important result of ARPA/SUR
1	7	8	Important
	3	3	Not important
			Other:

Comments: The future applicability of the systems is unclear; the technology was more important than meeting specs;

5. In addition to the final demonstrated versions of speech understanding systems, there was considerable work on advancing the methods of understanding and the effectiveness of various components or knowledge sources. Check whether, in comparison to the final system accuracy and specs, you think this advancement in specific aspects of recognition was:

6	9	15	Much more important
4	4	8	Somewhat more important
6	3	9	Equal in importance
	1	1	Less important
			Much less important

Comments: In the long run, the methods and components are much more important

Q5

6. On the left below, give a rank order of the fields of previous research and technology that were most effectively used in ARPA/SUR (1 = best used of all; 2 = next, etc.)

A	N	Best use of previous work T	FIELD	Most significant new contributions T	A	N
6.13	7.4	6.64	Acoustic Parameterization	4.04	3.00	5.25
3.33	5.0	4.00	Segmentation and Labelling	6.23	5.14	7.5
3.73	3.9	3.80	Phonological Rules	5.19	5.43	4.92
0.53	2.5	1.32	Prosodies	3.27	2.64	4.0
1.47	4.4	2.72	Word Matching and Verification	5.42	6.14	4.58
9.27	7.2	8.44	Syntax and Parsing	5.27	4.64	6.0
3.53	3.7	3.60	Semantics	3.00	2.0	4.178
1.06	3.6	1.68	Scoring Metrics and Procedures	3.08	3.57	2.5
3.40	3.0	3.24	Control Strategy, System Design	6.42	7.00	5.75
0.53	1.7	1.00	Performance Evaluation	1.15	1.5	0.75
0.60	0.0	0.36	Applications	0.35	0.64	0.0
			Research on			
			Other:			
			Other:			
			Other:			

7. On the right side of the above list, rank the fields in which you believe ARPA/SUR contributed the most significant new work or ideas.

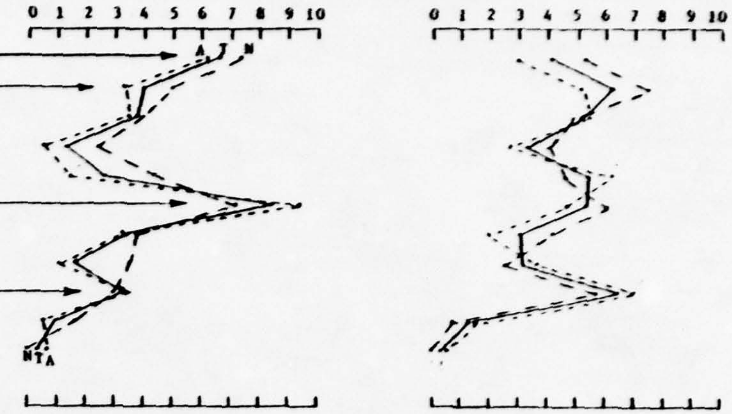
8. Some of the best technical contributions of ARPA/SUR are: (Make your own list)

See Section B-3.3

6&7. Average scores (rank 1 = score 10; rank 2 = 9, etc.) are plotted for ARPA (A), non-ARPA (N) and all (T) participants.

BEST USE OF PREVIOUS WORK

SIGNIFICANT ARPA SUR CONTRIBUTIONS



6. Both groups of respondents considered syntax and parsing, and acoustic parameterization, the top two best-used fields of previous work. Among the least-used were applications, performance evaluation, prosodics, and scoring techniques.

7. The fields of top-ranked contributions included control strategy and system design; segmentation and labeling, word matching and verification, syntax and parsing, and phonological rules. Other areas in which contributions substantially exceeded use of previous work were prosodics and scoring techniques.

8. Specifically mentioned contributions included: 21 mentions of the general area of control strategy and design of multiple-knowledge-source systems (concepts for organizing knowledge sources (8), control strategies (4), Harpy and the net matching scheme (4), and system organization (3)); eight (8) mentions of phonological rules (phonological rules (5), word boundary effects in word matching (2), and within- and across- syllable rules (1)); seven (7) mentions of acoustic phonetics (LPC uses and refinements (3), and one each for segmentation and labelling, acoustic phonetics, formant tracking, and detection of syllables and nasals); five (5) mentions of parsing and higher level linguistics (parsing in uncertainty, direction-free parsing, higher-level linguistic processing, analytic syntax, and parsing); and four (4) mentions of prosodics.

9. The Support Contractors of ARPA/SUR conducted basic and applied research in speech sciences and attempted to relate that work to the systems being developed. Answer each of the following questions with the one answer that best describes your assessment of their role:

a. The Support Contractors' work in the ARPA/SUR project was

	2	1	0	-1	-2	
	Very Important	Important	Somewhat Important	Of Little Importance	Of No Importance	
A+N - T:	1/2: 3	5/8: 13	9/5: 14	0/2: 2		0.53

b. The idea of independent research efforts was a good one.

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	
2/3: 5	12/10: 22	2/3: 5	0/1: 1		0.94

c. The system builders should have done this research themselves.

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	
0/1: 1	3/4: 7	6/5: 17	6/6: 12	1/0: 1	-0.16

d. The topics for the support studies were among the most needful of study.

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	
0/1: 1	4/5: 9	9/9: 18	2/0: 2	0/1: 1	0.22

e. The support efforts should have been directed primarily towards:

7/6: 13 Basic research relevant to SU

6/5: 11 Development of new knowledge sources or recognition methods

5/7: 12 Helping on problems defined by the system builders

No significant preferences shown.

Comments regarding the supporting research efforts:

See section B-3.4

-7-
Q7

10. How much has the ARPA SUR project advanced the state of the art in speech understanding from 1971 to now?

A/N T:

7/1: 8 A major breakthrough 6/7: 13 A significant advancement

2/6: 8 Some advancement 0/2: 2 Little advancement

0/1: 1 No significant effect

Comments: The net result may be summarized as a significant advancement bordering on a breakthrough.

11. Rate each of the ARPA SUR final-demonstration systems, by selecting the description that best fits your evaluation of each system (Mark one check in each column, at the appropriate answer)

A/N T: HARPY	HEARSAY II	BBN	SDC	
0/6: 6	3/8: 11	0/6: 6	3/13: 16	Don't know enough to evaluate
6/0: 6	2/0: 2	2/0: 2	0/0: 0	Excellent system, widespread utility
1/2: 3	2/1: 3	1/2: 3	0/1: 1	Excellent system, limited applicability
6/7: 13	6/6: 12	3/1: 4	1/0: 1	Good system but with limitations (describe below)
3/2: 5	1/0: 1	1/0: 1	0/0: 0	Good results, but wrong method (explain below)
0/0: 0	0/0: 1	0/1: 1	6/1: 7	Wrong design (explain below)
0/0: 0	0/0: 0	6/3: 9	2/1: 3	Poor results, but good design (explain below)
0/0: 0	0/0: 0	2/1: 3	3/0: 3	Other:
	1/0: 1			Other:

Further comments and explanations:

HARPY See section B-3.5

HEARSAY II

BBN

SDC

Q8

12a. Given one more year at similar funding levels, would the BBN system have met the original ARPA specs:

A/N I

5/2: 7 yes 3/1: 4 no 7/12: 19 uncertain.

Explain your decision: ____

12b. How about the SDC system:

0 yes 6/1: 7 no 8/13: 21 uncertain.

Explain: ____

12c. How about an SRI-developed system:

____ yes 4/1: 5 no 11/13: 24 uncertain

Explain: ____

The only yes votes were for the BBN system, and it had the fewest no votes, but the majority were uncertain what one more year would bring for any of the systems.

See section B-3.6 for comments.

13. A language is said to be "habitable" if users can readily learn it and keep within its constraints in actual conversations, so that spoken sentences are grammatical and otherwise correct and recognizable.

A/N

a. Do you consider the question of language habitability for SUS's to be 13 Very important 15 Important 3 Of interest 1 Of little importance 0 Of no importance

6/7 9/6 ; 1/2 0/1 0/0 — Quite important to both groups

b. Would you say habitability for forseeable SUS's is 0 Impossible to ever attain. 0 Impossible to obtain in the near future. 10 Very difficult to attain in a moderate size effort. 9 Difficult to attain. 9 Now within the scope of current capabilities. 1 Easy to attain. 1 Other: ____

0/0 0/0 ; 5/5 ; 5/4 5/4 ; 0/1 ; 1/0 — Difficult to attain

c. In answering (b), how were you considering the habitable languages to compare with those used in the ARPA SUR project? Habitable languages are

3/4 7 Far beyond the best language capability of ARPA SUR

4/5 9 Somewhat higher than the best ARPA SUR capability — Higher than the best ARPA SUR capability

5/3 8 Comparable to a moderate ARPA SUR-like language capability

2/ 2 Other: ____

d. Indicate for each of the following system's language-handling capabilities, how habitable you think they were

	Habitable	Almost habitable	Far from habitable	
CMU HARPY:	2	6	13	Not very close to being habitable, or "fairly far from--"
A/N:	1/1	4/2	9/4	
CMR HEARSAY II	2	6	10	
A/N:	2/0	3/3	7/3	
BBN:	6	8	8	Closest to being habitable
A/N:	5/1	4/4	5/3	
SDC:	0	5	10	
	0/0	4/1	5/5	
OTHER SYSTEM: ____				

Comments: See section B-3.7

PART C: BEST CURRENT TECHNIQUES

1. Rank the following systems in terms of their relevance to the way future work in SU should be done: (1= most relevant, 2= next, etc.)

A	N		
1.27	1.86	1.45	HEARSAY I
6.00	4.14	5.41	HEARSAY II
7.00	5.57	6.55	HARPY
1.47	1.29	1.47	DRAGON
2.53	1.43	2.18	SPEECHLIS
5.53	2.71	4.64	HWIM
1.67	-0.14	1.09	SDC 1973 VDM
1.33	1.29	1.32	SDC 1976 Demo System
2.93	1.57	2.50	SRI System
3.33	0.00	2.27	Lincoln Lab's System
2.20	1.86	2.09	UNIVAC System
4.47	5.14	4.68	IBM System
1.13	2.00	1.41	Bell Lab's System(s)

_____ Other Isolated Word Recognizers: TTI (1.43) SRI-SDC Demo System (0.67)

_____ Other Interstate (1.29) Locust System (CMU; o.53)

_____ Other Centigram (1.14)

Explain your choices briefly

See section B-4.1

Rating of 1 was assigned a score of 10, 2 assigned 9, 3 assigned 8, etc., then averages were compiled for the number of replies from each group. (Very low rating could thus come out negative, since 13 alternative systems were listed.)

2. Which search strategy is best in a speech understanding system:

A	N		
0	3	3	Best path first, with backtracking
3	5	8	Best few paths first, with backtracking
3	1	4	Best few paths first, no backtracking
0	0	0	Breadth first
1	1	2	Optimal test of all alternatives
2	2	4	Other:

Processing the best few paths first, preferrably with backtracting, is recommended. Breadth first is not favored at all.

3. "Island driving" and the use of "islands of reliability" is:

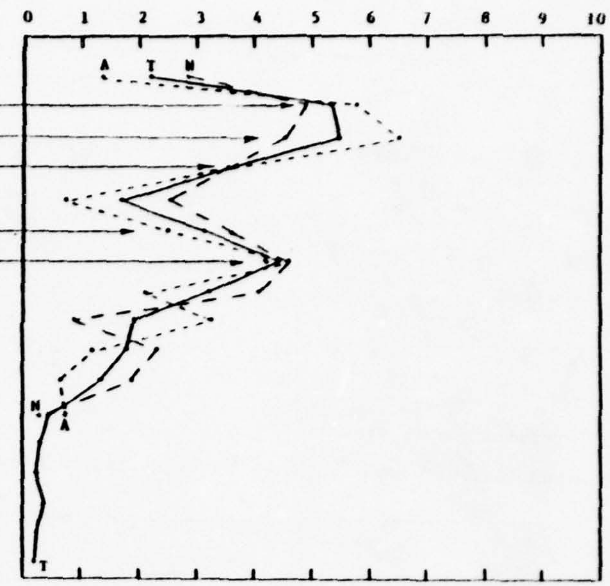
A	N		
0	1	1	A must in system design
1	3	4	Very valuable for system strategy
6	6	12	Useful provided the islands are truly reliable
3	3	6	Somewhat useful
3	0	3	Of limited value
0	0	0	Of no value

Islands of reliability are useful provided they are truly reliable.

Comments:

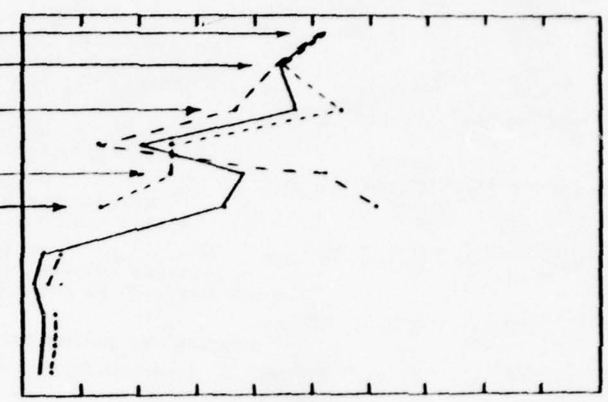
4. If work were to be done on new or dramatically improved knowledge sources or SUS components, to which would you most believe attention should be given? Rank order them (1_T = most, 2 = next, etc.):

A	N	T	
1.36	2.88	2.19	New acoustic parameters; namely, ________
5.71	4.94	5.29	Segmentation and labelling methods
6.50	4.53	5.42	Handling coarticulation effects
3.57	3.59	3.58	Phonological rules
0.71	2.59	1.74	Statistics of English
2.50	3.59	3.10	Prosodics
4.57	4.35	4.45	Word boundary effects and rules
2.07	4.06	3.16	Word matching procedures
3.21	0.94	1.97	Word verifier
1.21	2.35	1.84	Syntax and parsing
0.64	1.94	1.35	Semantics
0.71	0.29	0.48	Pragmatics
	0.53	0.29	Other: Phrase effects
	0.59	0.26	Other: Syllable effects
0.71		0.32	Other: Speaker normalization
0.57		0.26	Other: System Control and organization
0.50		0.23	Other: Automatic Lexical learning of pronunciations



5. If work were to be done on SU system structures for the "next generation" of systems, which methods would you choose first (rank order):

A	N	T	
5.19	5.15	5.17	An extended version of a HARPY-like system
4.50	4.38	4.45	A HEARSAY-like system of independent cooperating knowledge sources
5.50	3.77	4.72	A BBN-HWIM-like system
2.56	1.38	2.03	An SDC-SRI-like system
2.56	5.31	3.79	An IBM-like speech recognition system
1.38	6.08	3.48	A "Prosodically-guided" system (a la Lea, Medress, Skinner, 1975)
	0.77	0.34	Other: Threshold-like system
	0.38	0.17	Other: Entropy-guided system
0.56		0.31	Other: Original SDC system
0.56		0.31	Other: Locust
0.50		0.29	Other: Griffin (DB)



If you have recently been involved in the development of a speech recognition or understanding system, please list what you think are several of its most outstanding new ideas or contributions to the technology:

____

____

____

____

____

____

After evaluating your system, and seeing some of the interesting ideas and results of other groups, which of their ideas would you now include in your approach, if time and resources permitted:

____

____

____

____

____

10. The basic dictionary or lexicon in a speech understanding system should: A: 13 replies N: 14 replies

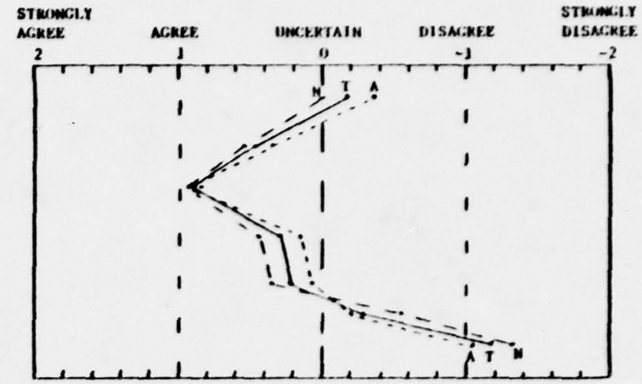
A/N T	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
a. Have only one entry per word	0/0: 0	0/1: 1	0/2: 2	8/8: 16	3/3
b. Have baseforms stated in phonemic forms	3/2: 5	6/5: 11	1/5: 6	0/2: 2	1/0: 1
c. Be fully expanded so no rules are necessary to specify expected pronunciations	3/0: 3	3/2: 5	2/5: 7	3/4: 7	0/2: 2
d. Have *a priori* probabilities assigned to alternative pronunciations	4/1: 5	3/5: 8	5/4: 9	1/3: 4	0/0: 0
e. Be a "syllabary" of expected pronunciations of syllables (not words)	1/1: 2	2/5: 7	5/5: 10	4/3: 7	0/0: 0
f. Consist of word 'templates' of acoustic (e.g., spectral) data, to be directly compared with unsegmented word-length segments in speech	1/2: 3	1/5: 6	3/2: 5	7/2: 9	1/2: 3
Other ______________					

11. Word verification in a speech understanding system is:

A/N T	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
a. Useful	10/6: 16	5/4: 9	0/1: 1	0	0
b. Currently quite effective	2/1: 3	5/1: 6	3/5: 8	4/2: 6	0
c. Best done with comparisons at a phonetic level	0/1: 1	3/5: 8	5/2: 7	4/0: 4	1/0: 1
d. Best done with comparisons at an acoustic level	4/0: 4	4/5: 9	3/2: 5	2/2: 4	0
e. Best done for every word at every possible point in the utterance	0	1/0: 1	3/4: 7	6/3: 9	4/2: 6
f. Best done only when the word is hypothesized with highscore	0/1: 1	6/3: 9	4/2: 6	4/2: 6	0
g. Other: ______________ 1/2					

8. Segmentation and labelling in a speech understanding system should be based on

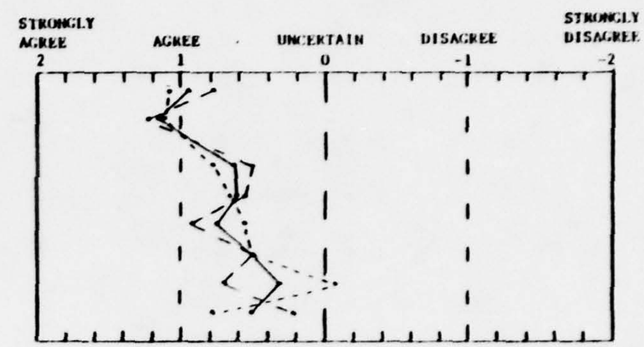
A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	A	N	T
Individual (10 ms) time samples -0.19	0/0: 0	3/5: 8	2/4: 6	4/3: 7	2/1: 3	-0.36	0.00	-0.19
A large set of alternative phonetic segments (a la HARPY)	2/2: 4	3/6: 9	4/3: 7	2/1: 3	0/1: 1	+0.36	0.54	+0.44
A lattice of phonetic alternatives	4/2: 6	6/7: 13	1/4: 5	1/0: 1	0/0: 0	+0.93	0.85	0.89
English phonemes (or something similar)	0/3: 3	5/4: 9	2/1: 3	1/4: 5	1/0: 1	+0.14	0.46	0.30
Diphones (from steady state of one phone to steady state of next)	0/1: 1	2/5: 7	6/5: 11	1/0: 1	0/1: 1	+0.07	0.38	0.22
Syllables as unanalyzed units	1/0: 1	1/3: 4	5/4: 9	6/6: 12	0/2: 2	-0.21	-0.54	-0.29
No segmentation below word level	0/0: 0	1/0: 1	0/2: 2	4/2: 6	6/8: 14	-1.07	-1.31	-1.18
Other: N: Phrase effects								
N: Hierirchy of segments, subphone-mic to syllable	3/1	0/1	0/0	0/0	0/0			



121

9. The following prosodic information should be extracted and used in a speech understanding system:

A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	A	N	T
Number of syllables	5/2: 7	5/8: 13	4/2: 6	0/1: 1	0/0	1.07	0.79	0.93
Locations of stressed syllables	4/6: 10	8/6: 14	2/2: 4	0/1: 1	0/0	1.14	1.21	1.18
Detection of intonational phrase boundaries	1/3: 4	9/4: 13	4/5: 9	0/1: 1	0/1: 1	0.79	0.50	0.65
Phrase final lengthening	2/3: 5	5/4: 9	7/6: 13	0/2: 2	0/0	0.64	0.57	0.61
Syntactic pauses	1/3: 4	8/8: 16	3/2: 5	2/1: 3	0/0	0.57	0.93	0.75
Intonation contours	1/2: 3	5/6: 11	8/4: 12	0/1: 1	0/1: 1	0.50	0.50	0.50
Rhythm and interruptions	1/3: 4	2/6: 8	7/3: 10	3/2: 5	1/0: 1	-0.07	0.71	0.32
Rate of speech	2/2: 4	8/3: 11	3/5: 8	1/2: 3	0/1: 1	0.79	0.21	0.50
Other:	0/1: 1							



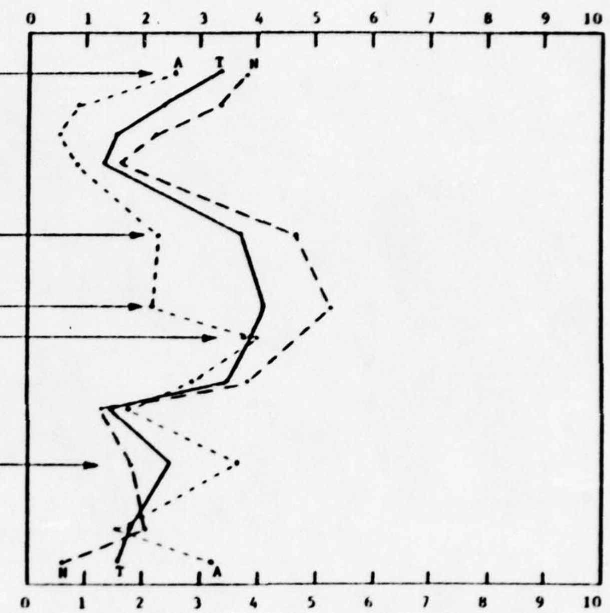
12. Syntactic analysis in a speech understanding system should:

A+N - T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	A	+ N	- T
a. Sharply constrain ambiguities in wording	2/1: 3	4/3: 7	1/3: 4	2/2: 4	0	0.05	0.27	0.41
b. Be combined with semantics, in a "pragmatic grammar"	1/1: 2	6/5: 11	2/3: 5	0/1: 1	0	0.73	0.55	0.64
c. Use an augmented transition network (ATN) grammar	2/0: 2	1/3: 4	4/5: 9	1/1: 2	0	0.36	0.18	0.27
d. Control the system in a "topdown" manner	2/2: 4	3/2: 5	1/2: 3	1/2: 3	2/0: 2	0.18	0.36	0.27
e. Confine the language to finite state form, without extensive context dependencies	1/0: 1	1/2: 3	4/1: 5	3/4: 7	0/1: 1	0.0	-0.36	-0.18
f. Use an "island-drive" strategy	1/0: 1	2/5: 7	2/4: 6	3/1: 4	0	0.09	0.36	0.23
g. Use a "best-first" parsing (search) strategy	0/0: 0	1/1: 2	6/5: 11	2/2: 4	0	-0.09	-0.09	-0.09
h. Use a "depth-first" parsing (search) strategy	2/0: 2	0/1: 1	5/7: 12	1/0: 1	1/0: 1	0.09	0.09	0.09
i. Use a "best-few-first", or beam search strategy	3/0: 3	3/6: 9	3/3: 6	1/1: 2	0	0.73	0.45	0.59
j. Other: ______	0/1: 1							

PART D: CURRENT NEEDS AND GAPS IN SPEECH RECOGNITION TECHNOLOGY

1. Rank order the types of speech recognition or understanding that are <u>most needed now</u>:

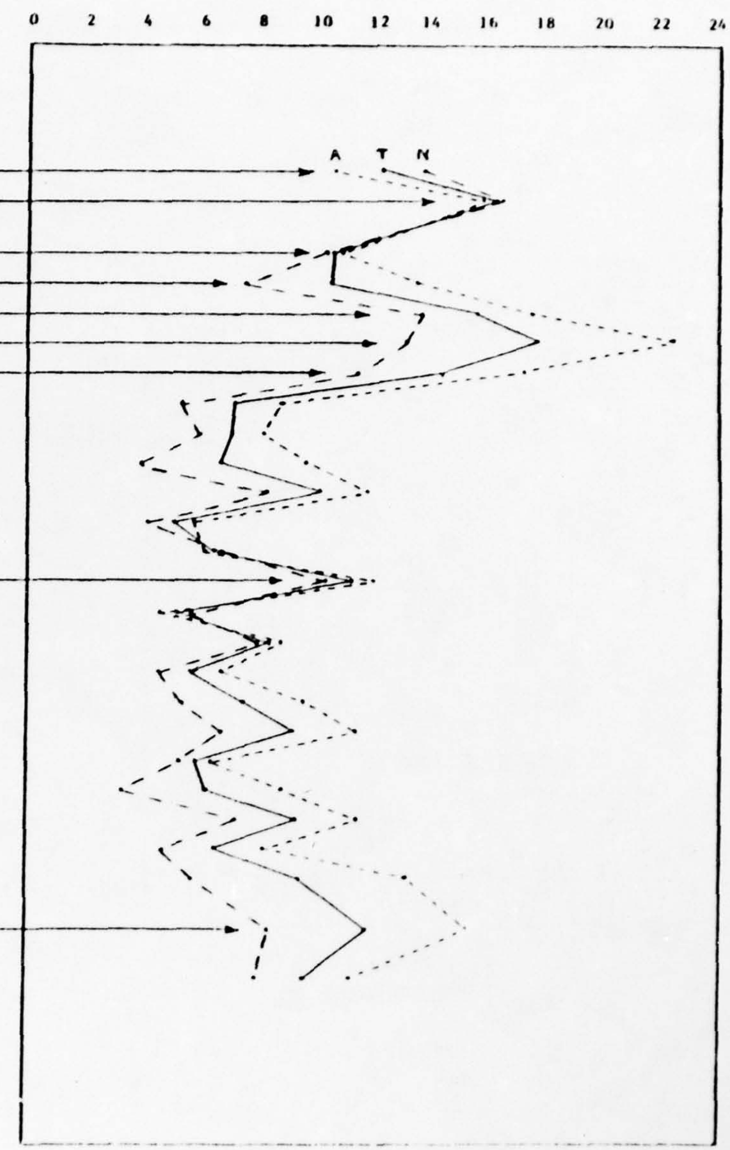
A	N	T	
2.55	3.88	3.36	Very low cost ($______) isolated word recognizers
0.91	3.35	2.39	Moderate cost ($______) isolated word recognizers
0.55	2.24	1.57	Larger-vocabulary (_____-word) isolated word recognizers
0.91	1.59	1.32	Recognizers that handle sequences of words, where successive words are separated by short pauses, without use of higher-level constraints
2.27	4.71	3.75	Recognizers that handle sequences of words separated by short pauses, with use of higher-level (syntactic, semantic) constraints
2.18	5.36	4.11	Digit string recognizers
4.0	3.88	3.91	Recognizers of connected word sequences that are highly constrained to certain formats
2.91	3.88	3.50	Recognizers of modestly-constrained word sequences
1.82	1.29	1.47	Speech understanding systems that are much more restricted than HARPY
3.64	1.76	2.50	Fairly-well-restricted (HARPY-like) sentence understanding systems
1.55	2.0	1.82	Speech understanding systems that are somewhat less restricted than HARPY
3.18	0.53	1.57	Speech understanding systems that are much more powerful (unrestricted) than HARPY
		_____	Other: ____________________
		_____	Other: ____________________
		_____	Other: ____________________

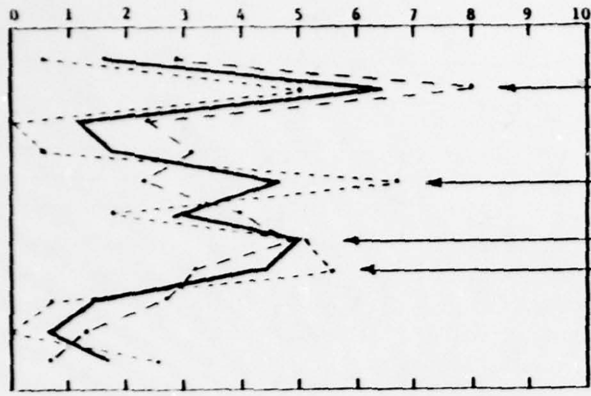


2. What are the biggest (or most troublesome) "bottlenecks" in speech understanding technology today (rank order in decreasing importance: 1 = most important, 2 = next, etc., or else mark each factor as of High, Medium, or Low importance)? That is, which of the following need the most attention? Give any specifics about each need, or group of needs that you can, and attach extra sheets or relevant information if you desire:

H=25 points, M=15 points, L=5 points, Rank1=30 points, 2=29 points, 3=28 points

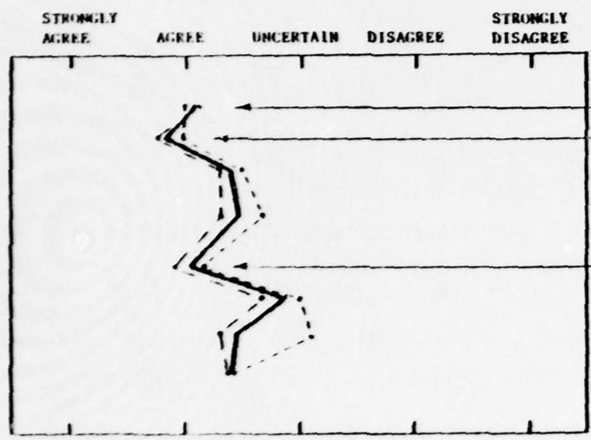
Rank:	Need:	Comments: A	N	T
5	Acoustic parameterization techniques	10.57	13.86	12.22
2	Segmentation procedures	15.86	16.36	16.11
8	Sizes of segments (acoustic, phonetic, diphonic, syllabic, etc.	10.93	10.29	10.61
9	Phonetic or spectral distance measures	13.50	7.57	10.54
3	Labelling of segments	17.43	13.79	15.61
1	Acoustic phonetic rules	22.29	13.07	17.68
4	Phonological rules	17.07	11.50	14.29
17	Dictionary entries	8.86	5.36	7.11
18	Dictionary expansions by rules	8.07	6.00	7.04
19	Lexical retrieval	9.64	4.00	6.82
10	Word matching procedures	11.86	8.21	10.04
26	Scoring procedures	5.86	4.21	5.14
21	Word verification	6.14	6.79	6.47
7	Prosodic aids to parsing	12.00	10.29	11.15
25	Language statistics	4.64	5.64	5.14
15	Use of syntactic constraints	8.79	8.00	8.40
24	Parsing procedures	6.79	4.71	5.75
16	Habitability of language	9.57	5.43	7.50
14	Use of semantic constraints	11.36	6.86	9.11
23	Selection of task domains	6.43	5.43	5.93
22	User models	8.93	3.29	6.11
13	System control and architecture	11.43	7.21	9.32
20	Measures of task complexity	8.21	4.86	6.54
12	System (and component) performance metrics	13.07	5.86	9.47
6	Amount of data processed through system during design	15.21	8.43	11.82
11	Research on acoustic phonetic characteristics of English sentences	11.21	8.00	9.61
(1)	Research on Word boundary detection; 2: Res. on sentence boundary detection			
(1)	Other: Project organization, management, and maintenance of large system; 2: Response Time of system.			
H	Other: Speaker Independence H: Research on ideolectal, dialectal, & other sources of variability.			
H	Other: Telephone Input; H: practicality. no existing capability is reduced to microsystems.			
H	Other: Ways for human to learn from analysis of lots of data, with minimal effort spent hand labelling speech.			





3. Rank order the following sets of acoustic parameters in terms of their utility in speech understanding

A	N		
0.55	2.9	1.67	zero crossing counts, in _____ bands
5.00	8.0	6.43	3 (or ___) formant frequencies
0.00	2.4	1.14	formant bandwidths
0.55	3.1	1.76	formant amplitudes
6.73	2.3	4.62	linear-predictive analysis
1.82	4.3	3.00	poles of linear predictive spectrum
5.09	4.9	5.0	fundamental frequency
5.55	3.2	4.43	amplitude or energy measure
0.73	2.8	1.48	time domain analysis features, namely, __________
0.0	1.4	0.67	instantaneous frequency (a la Janet Baker, CMU)
2.55	0.7	1.67	total frequency spectrum (for direct comparison with templates or expected spectra)
		____	Other: __________
		____	Other: __________
		____	Other: __________



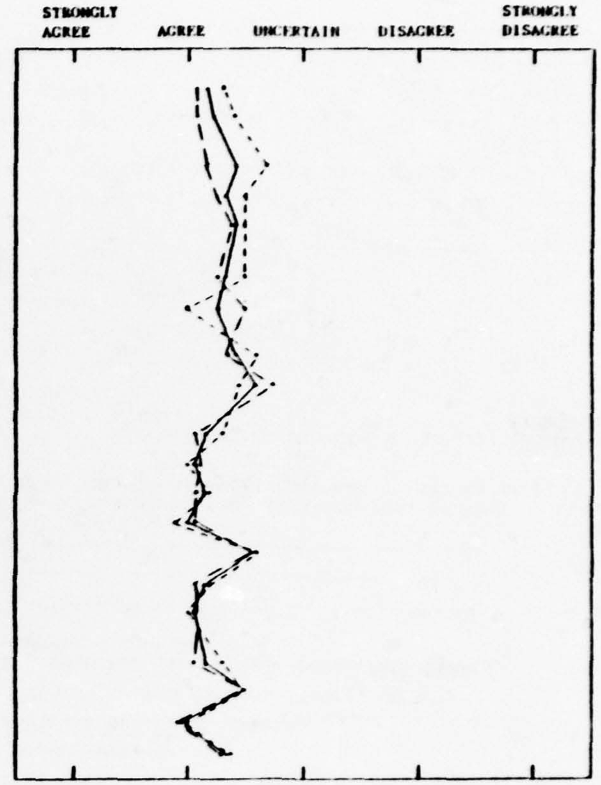
4. For future speech recognition/understanding studies, further work is needed on the following acoustic parameter extraction processes:

A	N	T	A/N T:	Strongly Agree +2	Agree +1	Uncertain 0	Disagree -1	Strongly Disagree -2
1.00	0.85	0.92	Selection of Acoustic Parameters	5/5: 10	4/7: 11	0/1: 1	2/1: 3	____
1.00	1.23	1.12	Distance Measures	4/4: 8	5/8: 13	1/2: 3	1/0: 1	____
0.50	0.69	0.60	Formant Tracking Procedures	1/1: 2	6/8: 14	3/4: 7	2/1: 3	____
0.33	0.69	0.52	Vowel Formant Target Frequencies	1/1: 2	4/8: 12	5/4: 9	2/1: 3	____
0.83	1.08	0.96	Vowel Formant Transitions and Contours	3/3: 6	5/8: 13	3/3: 6	1/0: 1	____
0.0	0.31	0.16	Use of LPC Poles	1/0: 1	2/7: 9	5/4: 9	4/1: 5	0/1: 1
-0.17	0.69	0.54	Fundamental Frequency Tracking	0/1: 1	2/9: 11	6/3: 9	4/0: 4	0/1: 1
0.58	0.62	0.60	Voicing Decisions	2/1: 3	4/8: 12	5/4: 9	1/0: 1	0/1: 1
			Other: __________	____	____	____	____	____
			Other: __________	____	____	____	____	____
			Other: __________	____	____	____	____	____

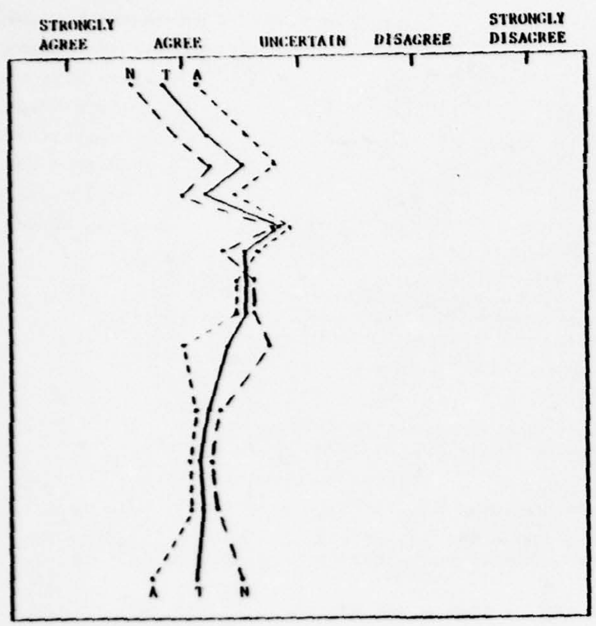
Note: "Already acceptable" pooled with "strongly disagree" in calcuting mean positions.

5. For future speech recognition/understanding studies, further work is needed on the following segmentation and labelling procedures:

A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	Already Acceptable	A	N	T
Locating syllabic nuclei	2/3: 5	5/6: 11	1/3: 4	0/1: 1			0.7	0.92	0.82
Locating syllable boundaries	2/3: 5	3/6: 9	3/3: 6	1/1: 2			0.6	0.92	0.77
Segmenting and identifying diphones	1/1: 2	2/8: 10	2/3: 7	1/0: 1			0.3	0.83	0.59
Detecting vowels	2/3: 5	4/6: 10	2/2: 4	1/1: 2		/1 1	0.5	0.75	0.64
Determining vowel height	1/2: 3	3/5: 8	5/3: 8	0/2: 2			0.5	0.58	0.55
Determining vowel place (front/back, etc.)	1/2: 3	3/6: 9	5/3: 8	0/1: 1			0.5	0.75	0.64
Detecting diphthongs	2/3: 5	7/3: 10	0/4: 4	1/1: 2		/1 1	1.0	0.5	0.73
Distinguishing among diphthongs	1/3: 4	5/5: 10	1/4: 5	1/0: 1			0.4	0.67	0.55
Detecting retroflexives	2/1: 3	4/4: 8	3/5: 8	0/1: 1		/1 1	0.6	0.25	0.41
Detecting and identifying glides	2/3: 5	5/6: 11	1/2: 3	0/1: 1			0.7	0.92	0.82
Detecting laterals	3/2: 5	4/5: 9	2/4: 6	0/1: 1			1.0	0.83	0.91
Detecting nasals	2/3: 5	6/7: 13	1/1: 2	0/2: 2			0.8	0.92	0.86
Distinguishing among nasals	3/3: 6	5/7: 12	1/1: 2	0/0: 0		0/1: 1	1.1	0.92	1.00
Detecting sibilants	1/2: 3	5/5: 10	2/2: 4	1/2: 3		/1: 1	0.4	0.42	0.41
Distinguishing among sibilants	1/4: 5	5/5: 10	2/2: 4	1/0: 1		/1: 1	0.8	0.92	0.86
Detecting other fricatives	2/2: 4	6/7: 13	2/2: 4	0/0: 0			1.0	0.92	0.95
Distinguishing among fricatives	1/2: 3	5/8: 13	2/1: 3	0/1: 1			0.7	0.92	0.82
Detecting stop consonants	3/2: 5	2/6: 8	2/1: 3	1/2: 3		/1: 1	0.5	0.5	0.50
Distinguishing among stops	2/3: 5	6/7: 13	1/2: 3	0/0: 0			1.0	1.08	1.05
Detecting affricates	2/0: 2	4/9: 13	2/1: 3	0/1: 1			0.6	0.67	0.64
Other:	0/2: 2								
Other:	0/1: 1								
Other:									



Comments



6. To more fully characterize your assessment of speech recognition needs, without the detailed specifics in a previous question, but with more chance to show the degree of your convictions, please answer the following question.

The following are among the most significant "gaps"(problem areas in need of further work) in speech understanding technology

A	N	T	A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
0.87	1.44	1.16	a) Acoustic phonetic analysis (the "frond end" of a system)	10/10: 20	4/4: 8	0/2: 2	1/1: 2	0
0.47	1.06	0.77	b) Prosodic cues to linguistic structures	2/6: 8	5/5: 10	4/4: 8	2/0: 2	0
0.20	0.75	0.48	c) Phonological rules	0/2: 2	5/8: 13	6/5: 11	2/0: 2	0
0.53	1.0	0.77	d) Word verification	1/6: 7	7/8: 15	3/3: 6	1/2: 3	0/1: 1
0.07	0.19	0.13	e) Syntactic analysis	0/0: 0	4/8: 12	5/2· 7	3/3: 6	0/1: 1
0.33	0.63	0.48	f) Semantic analysis	1/1: 2	5/9: 13	6/4: 10	2/1: 3	0
0.53	0.38	0.45	g) Pragmatic analysis	1/1: 2	8/6: 14	3/6: 9	2/2: 4	0
0.53	0.38	0.45	h) Control strategy	2/3: 5	6/4: 10	2/5: 7	2/2: 4	0/1: 1
1.00	0.25	0.61	i) Scoring procedures	6/1: 7	4/5: 9	2/6: 8	1/1: 2	0/1: 1
0.87	0.69	0.77	j) Effective use of higher-level linguistic information to constrain ambiguities	4/3: 7	7/6: 13	1/6: 7	2/1: 3	0
0.93	0.75	0.84	k) Measures and methods for performance evaluation	5/4: 9	6/6: 12	0/3: 3	2/2: 4	0
0.93	0.69	0.81	l) Fast or near-real-time operation	5/7: 12	6/2: 8	2/2: 4	2/3: 5	0/1: 0
1.27	0.50	0.87	m) System "tuning" or adjustment based on extensive data processed	7/4: 11	6/4: 10	1/5: 6	1/2: 3	0/1: 0

Comments:

PART E: APPLICATIONS FOR ASR/SUS

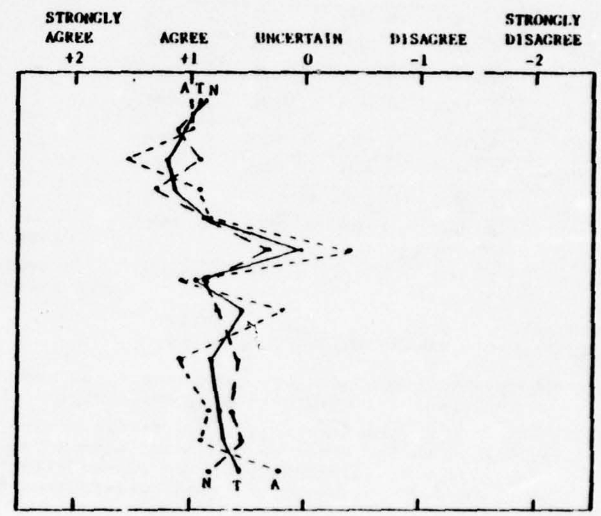
1. Some work is currently going on to develop SUS's for which the talker says a series of isolated words (separated by short but detectable pauses), and whole sentences are thus composed, sometimes with syntax and semantics helping disambiguate and reduce errors in the sequence. Would you say such a mode of man-computer interaction is:

A	N		
2	4	6	Clearly feasible for humans to use and machines to handle
6	5	11	Better for machines than connected speech, and only somewhat difficult for the human to use
6	6	12	Quite difficult for the human to use
0	0	0	Impossible for the human
2	5	7	In need of experimental study to determine its effectiveness, before/after (cross out one) development of the systems
		___	Other: ______

There is some uncertainty about just how difficult it may be for humans to use sequences of isolated words, but it seems to warrent further study.

2. Indicate whether you agree that the following specific applications of speech recognition technology are very important and appropriate for future work:

A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	A	N	T
Package Sorting	3/4: 7	7/9: 16	2/1: 3	0/2: 2	0	1.0	0.88	.93
Inventorying	3/6: 9	7/8: 15	3/1: 4	0/2: 2	0	1.0	1.06	1.03
Data Retrieval	7/4: 11	6/10: 16	0/0: 0	0/2: 2	0	1.54	0.94	1.20
Air Traffic Control	4/6: 10	4/10: 14	4/1: 5	0	0	0.92	1.29	1.13
Cartography	3/4: 7	5/8: 13	3/4: 7	0/1: 1	0	0.85	0.88	0.87
Gun fire control	1/1: 2	1/8: 9	6/5: 11	2/1: 3	3/1: 4	-0.38	0.35	0.03
Command and control systems	4/7: 11	6/8: 14	3/2: 5		0	1.08	0.88	0.97
Spotting key words in context	0/2: 3	6/10: 16	5/3: 8	1/1: 2	1/0: 1	0.23	0.76	0.53
Computer aided troubleshooting of equipment	5/4: 9	4/6: 10	3/3: 6	0/4: 4	0	1.08	0.59	0.80
Assisting computer graphics interactions	3/3: 6	5/7: 12	4/4: 8	0/2: 2	0	0.85	0.65	0.73
Airline reservations	5/3: 8	3/5: 8	3/5: 8	1/1: 2	0/1: 1	0.92	0.53	0.70
Voice Programming	1/5: 6	5/5: 10	3/4: 7	2/1: 3	1/0: 1	0.23	0.82	0.52
Other ______	0/1: 1	0/1: 1						



The exponential scale of alternative market sizes requires a logarithmic use of the individual judgements, so that by counting the number of judgments in each bin, and getting the mean position M by this equation for a total of N responses

$$M = (\# \text{ in 10-bin})(1) + (\# \text{ in 100-bin})(2) + (\# \text{ in 1000-bin})(3) + (\# \text{ in 10,000-bin})(4) + (\# \text{ in 100,000-bin})(5)$$

we get the logarithm of the average number of units, and then taking the antilogarithm yields these market estimates. This is calculated for each group of respondents:

A	N	T
3466	2290	2753
1513	371	692
813	513	631
5246	6164	5753
813	1949	1318
436	195	275
288	162	209
355	162	229
1000	195	398
436	851	631

Total Estimate per year: 12,889

3. Given the current state of technology in speech recognition and understanding, and whatever applications you are aware of, how large a market would you estimate there to be (now or in the next few years) for the following types of recognizers, in terms of systems to be bought per year:

Small vocabulary (10-50 word) isolated word recognizers

0/1: 1 None 0/0: 0 10 0/0: 0 100 1/4: 5 1000 4/5: 9 10,000 1/0: 1 100,000 0/1: 1 Other: 0

Larger vocabulary (_____-word) isolated word recognizers

0/1 ___ None 0/1: 1 10 2/3: 5 100 1/5: 6 1000 2/0: 2 10,000 1/1: 2 100,000 0 Other: 0

Recognizers of word sequences ("sentences") with short pauses between words

0/1: ___ None 0 10 1/3: 4 100 2/4: 6 1000 2/2: 4 10,000 0 100,000 0 Other: 0

Digit string recognizers

0 None 0 10 0 100 1/3: 4 1000 3/4: 7 10,000 1/2: 3 100,000 1/1: 2 Other: 0

Recognizers of connected word sequences, without pauses, but highly formatted

0 None 0 10 2/ :2 100 2/8: 10 1000 1/2: 3 10,000 0/1: 1 100,000 0 Other: 0

Highly restricted speech understanding systems (less than HARPY)

1/2: ___ None 0/1: 1 10 0/3: 3 100 2/1: 3 1000 1/1: 2 10,000 0 100,000 0 Other: 0

Moderate speech understanding systems (HARPY, or slight extensions)

1/2: ___ None 1/2: 3 10 1/2: 3 100 2/2: 4 1000 0/1: 1 10,000 0 100,000 0 Other: 0

Ambitious speech understanding systems (at highest ARPA SUR goals or somewhat higher)

1/2: ___ None 2/3: 5 10 0/1: 1 100 1/3: 4 1000 0/2: 2 10,000 1/ : 1 100,000 0 Other: 0

Fully versatile speech understanding systems for large habitable subsets of spoken English

2/3: ___ None 0/1: 1 10 0/2: 2 100 1/1: 2 1000 0/3: 3 10,000 3/0: 3 100,000 0 Other: 0

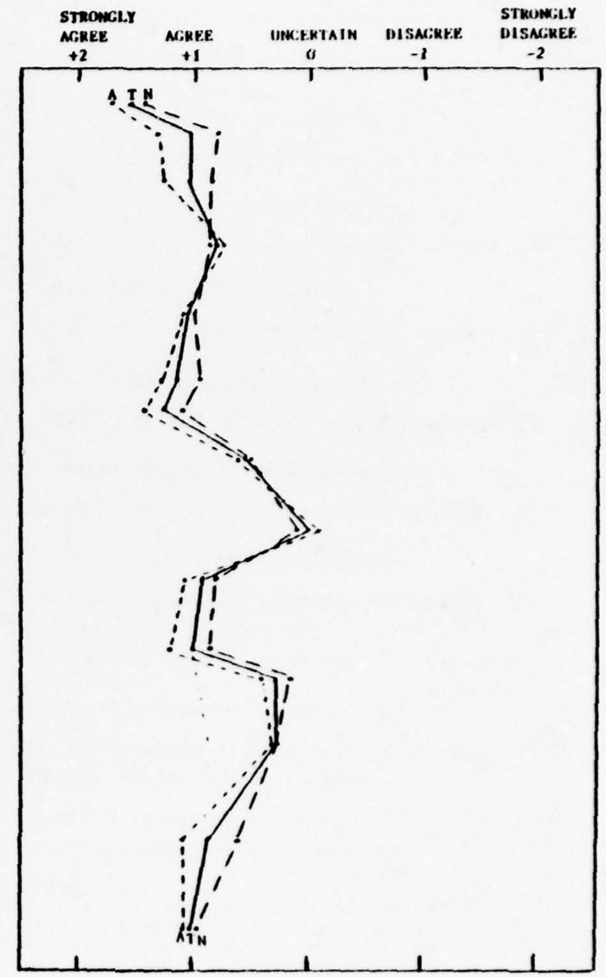
Connected speech recognition without task or language constraints (a la IBM)

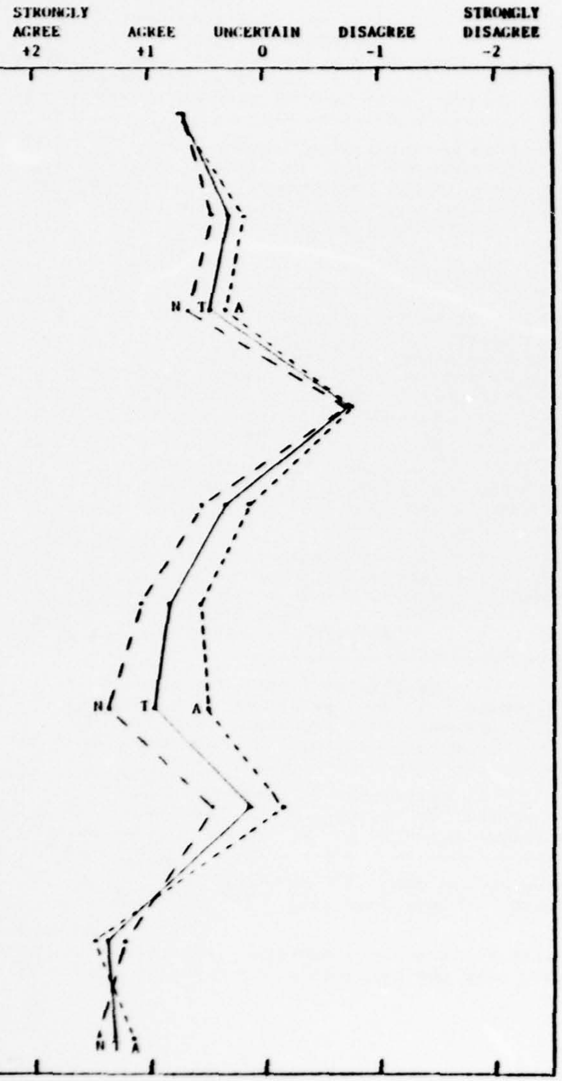
2/1: ___ None 0 10 2/1: 3 100 0/3: 3 1000 0/1: 1 10,000 2/1: 3 100,000 0 Other: 0

Others, or comments: ______________________________

4. Indicate which of the following advantages of speech are important in determining when and how speech recognition or understanding systems should be used. It is important that speech is:

A/N T:	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	A	N	T*
(a) Human's most natural communication modality	11/9: 20	4/8: 12	0/1: 1	0	0	1.73	1.44	1.58
(b) Easy to use without training	7/6: 13	6/8: 14	1/1: 2	0/1: 1	0	1.33	0.78	1.03
(c) Human's highest-capacity output channel	5/6: 11	9/6: 15	1/2: 3	0/3: 3	0	1.27	0.83	1.03
(d) Suitable for talking to humans and machines simultaneously	4/5: 9	5/7: 12	4/3: 7	2/2: 4	0	0.73	0.83	0.79
(e) Possible in the dark, at various orientations, and around corners, without visual contact	5/3: 8	7/12: 19	2/3: 5	1/0: 1	0	1.07	1.00	1.03
(f) Possible in conjunction with other modalities, for multi-modal communication	6/4: 12	7/9: 16	2/3: 5	0	0	1.27	0.94	1.12
(g) Possible while talker is mobile	8/5: 13	5/10: 15	3/3: 6	0	0	1.40	1.11	1.24
(h) Also useful for detecting a talker's identity	1/1: 2	9/8: 17	3/6: 9	2/1: 3	0	0.60	0.50	0.55
(i) Also useful for detecting a talker's physical and emotional state	1/1: 2	1/5: 6	10/5: 15	1/5: 6	2/ : 2	-0.13	0.11	0.00
(j) Easily transduced with lightweight portable microphones	4/3: 7	8/8: 16	2/6: 8	0	0	1.07	0.78	0.91
(k) Compatible with telephones, radios, and voice communication networks	6/3: 9	6/9: 15	3/4: 7	0	0	1.2	0.83	1.00
(l) Unaffected by weightlessness	2/1: 3	3/5: 8	8/6: 14	1/4: 5	0	0.40	0.17	0.27
(m) Only slightly affected by moderate g-forces (accelerations)	0/1: 1	5/5: 10	8/8: 16	1/2: 3	0	0.26	0.28	0.27
(n) Demonstrated to be the most effective single modality in *problem solving* tasks (Chapanis, et al.)	6/3: 9	4/6: 10	5/6: 11	0/1: 1	0	1.07	0.61	0.82
(o) Demonstrated to be an integral part of the most effective multimodel communication links (Chapanis, et al.)	6/5: 11	4/7: 11	4/4: 8		0	1.07	0.94	1.00
(p) Other:	1/1: 2							





PART F: FURTHER LONG-RANGE WORK

1. If another large-scale project in SUS were undertaken, I would include the following organizational features:

A	N	T	Item	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
			a. Construction of several alternative systems					
0.71	0.73	0.72		3/3: 6	9/9: 18	1/3: 4	1/0: 1	1/2: 3
			b. Completdly-defined system specs					
0.18	0.40	0.28		3/3: 6	4/6: 10	2/2: 4	5/6: 11	1/0: 1
			c. Strict deadlines for system demonstration					
0.29	0.67	0.47		2/3: 5	6/8: 14	2/3: 5	5/2: 7	0/1: 1
			d. Close attention to immediate military applications					
-0.76	-0.73	-0.75		0	2/4: 6	4/2: 6	3/7: 10	6/4: 10
			e. A Steering Committee for program management					
0.12	0.53	0.31		0/3: 3	8/5: 13	2/6: 8	4/3: 7	1/0: 1
			f. Support contractors to do related research					
0.53	1.07	0.78		1/2: 3	7/12: 19	7/3: 10	0	0
			g. Mid-term preliminary evaluations of systems					
0.47	1.33	0.88		3/5: 8	5/11: 16	4/0: 4	3/1: 4	0
			h. Elimination of poorest systems at mid-term					
-0.18	0.47	0.13		0/2: 2	3/5: 8	7/8: 15	4/2: 6	1/0: 1
			i. An extensive performance evaluation study following completion of system development. (Evaluation of all system components, strategies, etc.)					
1.47	1.2	1.34		10/6: 16	5/7: 12	0/2: 2	0/1: 1	0
			j. Close interactions and interchanges *among* contractors					
1.18	1.4	1.28		6/7: 13	8/8: 16	0/1: 1	0/1: 1	0

A/N T

2. If another large-scale project in SUS development were undertaken, I would include the following system design choices:

A/N I:

0/1 9/7 3/5 3/2 0/1 a. Vocabulary of 1 Less than 100 16 Several hundred 8 Several thousand 7 Tens of thousands of words 1 Other: ____

2/2 2/1 4/2 3/7 0/1 0/1 4/4 0/1 b. Speaker population 4 1 or a few selected talkers 3 Less than 10 6 10-50 10 50-100 1 100-1000 1 Many thousand 8 Unlimited 1 Other: ____

0/1 0/3 2/4 8/7 7/4 0/1 c. Form of speech 1 Isolated words 4 Digit strings 6 Strictly formatted short sentences 15 Sentences related to a restricted task 11 Almost any sentence in a versatile subset of English 1 Other: ____

d. Noise and channel conditions
6/3: 9 In a quiet room
9/6: 15 In a normal computer room (with some noise)
4/6: 10 Over a high-quality communication channel
3/7: 10 Over a noisy communication channel
2/1: 3 Other: ____

4/6 8/7 7/7 2/6 0/ e. Input conditions: 10 Good quality room microphone 15 Close-talking microphone 14 Telephone 8 Radio Channel 0 Other: ____

f. Tuning to speaker
4/4: 8 Totally speaker-independent
12/9: 21 Adjustable to speaker, using small set of utterances
0/5: 5 Speaker speaks most or all possible utterances once, and system adjusted on basis of them .
0/1: 1 Speaker speaks several repetitions of all possible utterances, and system uses those "templates".
0/0: 0 Other: ____

g. Use of syntax
1/1: 2 Don't use syntactic constraints
1/3: 4 Use syntactic constraints only slightly
7/12: 19 Make a fair amount of use of syntactic constraints
8/3: 11 Rely heavily on syntactic constraints
3/0: 3 Make syntactic constraints control system choices
____ Other: ____

h. Semantic support (use of task dependent knowledge to reduce search space)
10/11: 21 Make substantial use of semantic, pragmatic, and user model constraints
3/3: 6 Use semantics and discourse pragmatics, but no user model
1/1: 2 Use semantics only ____ Use discourse constraints only
1/2: 3 User model only ____ Don't rely on semantic and task constraints
0/2: 2 Ignore these
0/ ____ Other: ____

i. Real-time operation
1/6: 7 An absolute must
6/8: 14 Very desirable
2/3: 5 OK, if readily obtained
2/0: 2 Not important
0/1: 1 A diversion, don't waste time on it
0/ ____ Other: ____

j. Computer: Requiring the following computer
1/4: 5 Minicomputer (PDP-11 or less)
7/8: 15 PDP-10-size
2/1: 3 - Million instructions per second ____
1/0: 1 Multiprocessor, as follows: ____
0/2: 2 Other: interface with high-speed signal processor for speed

k. Accuracy required (semantically):
4/11 5/6 3/2 1/1: 15 Over 99% 11 Over 95% 5 Over 90% 2 Over 80%
0/0 ____ Other: ____

l. Time Scale
0/0: 0 1 yr. 0/3: 3 2 yr. 2/3: 5 3 yr. 5/2: 7 4 yr. 0/1: 1 5 yr
2/2: 4 10 yr.
0/3: 3 Other: ____

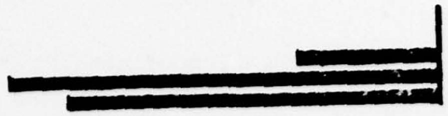
m. Types of Tasks
0/0: 0 Games (chess, etc.)
8/9: 12 Information Retrieval (specify type): ____
3/11: 14 Command and Control systems (specify type): ____
1/7: 8 Voice Programming (specify language, etc. ____
9/4: 13 Several different tasks, covering a range of difficulties
0/1: 1 Other: Allow task to be dynamically defined

N. Other Design

Choices (Specify): ____

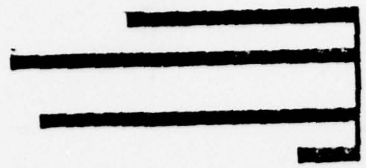
A/N T: 3. I would recommend that work on speech recognition or understanding systems be:

(a) The overall recommendation might be summarized as a call for a moderate increase in funding for speech recognition work.



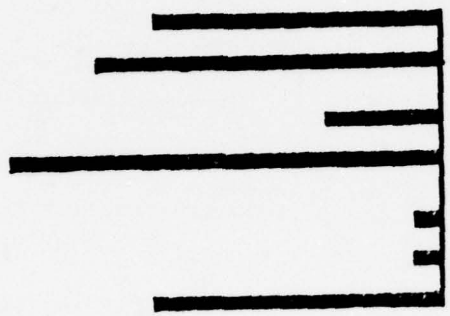
0/0: a. 0 Ended now
0/0: 0 Sharply curtailed
1/4: 5 Kept at about its current level of funding and effort
7/8: 15 Increased somewhat in funding and effort
7/6: 13 Stimulated by a large increase in funding and effort

(b) Respondents were fairly evenly split between distinct modest projects, a carefully coordinated set of modest projects, or a large coordinated project coupled with current modest projects. Over 2/3 are calling for some form of coordinated effort.



b. The work should be funded by:
3/5: 8 Supporting many separate modest size projects, for various applications, agencies, and distinct purposes
3/9: 12 Supporting many modest projects that are carefully coordinated. (What method of coordination? ______ ______)
8/3: 11 Current modest projects, plus a large coordinated project like ARPA SUR
1/1: 2 The following method(s): ______

(c) The effort should be predominantly research and development, with new techniques and systems being researched with applications in mind.



c. Recognized (and supported) as (currently and in the next few years) predominantly an effort in
4/6: 10 Long-range research on speech characteristics, for later applications
6/6: 12 Applied research, in which major questions of recognition techniques still need to be answered before applications are addressed
1/3: 4 Development, in which prototype systems are to be developed using available technology
6/9: 15 Applied research and development, in which recognition techniques are being developed and applied to develop prototype systems
0/1: 1 Producing copies and slight improvements of currently available prototype systems
0/1: 1 Production of off-the-shelf devices (or systems) in quantity
4/6: 10 A combination of all the above
____ Other: ______

PART G: REACTIONS AND FURTHER REMARKS

1. This survey was a good way to:

   a. Review the ARPA/SUR project

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
1/1: 2	5/7: 12	5/9: 14	2/1: 3	1/ : 1

   b. Survey the current work

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
0/1: 1	3/4: 7	6/9: 15	2/3: 5	3/ : 3

   c. Survey opinions about future ASR/SUS work

Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
2/3: 5	7/12: 19	4/2: 6	0	1/ : 1

   Comments: ______________________

2. I would recommend the following improvements in the survey: ______________________

3. Please send this survey questionnaire to the following qualified person(s): ______________________

4. We had the following general "essay" questions in mind when we prepared this questionnaire. *If we have missed* any aspects of them in what you have already replied to, or if they provoke other thoughts, please include your remarks in the space provided, or attach further information. Essay questions:

   - o What contributions came from the ARPA SUR project?
   - o What is the overall evaluation of the system performance?
   - o What problems in SU became evident during ARPA SUR?
   - o Which system components worked best?
   - o What are the primary sources of error and limitations in the ARPA systems?
   - o How "extendable" and "versatile" does a SUS have to be? Is HARPY good enough?
   - o How does one comparatively evaluate quite different systems such as the ARPA SUR ones?
   - o How do other systems *compare* with ARPA SUR results?
   - o What are appropriate task domains for speech understanding?
   - o How far are we from "habitable" languages in SUS?
   - o Are there any real applications for the various SU system capabilities?
   - o What are the best available techniques in ASR/SUR?
   - o What "gaps" remain in speech understanding technology?
   - o Is there any need for simulations (by human) of expected machine performances, before total systems have to be built?
   - o What research needs to be done?
   - o What aspects of system design deserve work now?
   - o How should further projects in SUS/ASR be conducted?
   - o Where do we go from here in speech recognition/understanding?

Comments: ______________________

(9) Final rept. 20 Jul 77-19 Jul 78,

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
(6) REVIEW OF THE ARPA SUR PROJECT AND SURVEY OF CURRENT TECHNOLOGY IN SPEECH UNDERSTANDING.

5. TYPE OF REPORT & PERIOD COVERED
Final Report
July 20, 1977-July 19, 1978

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
(10) Wayne A. Lea ~~and~~ June E. Shoup

8. CONTRACT OR GRANT NUMBER(s)
(15) N00014-77-C-0570
NR 049-413

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Speech Communications Research Laboratory
806 West Adams Boulevard
Los Angeles, CA 90007

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
(11) 16 Jan 79

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Department of Navy
806 North Quincy Street
Arlington, Virginia 22217

(12) 162p.

12. REPORT DATE
January 16, 1979

13. NUMBER OF PAGES
xxv + 134 = 159 total

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Unlimited

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech recognition	Phonological rules
Speech understanding	Syntactic analysis
Speech processing	Semantic analysis
Voice input to computers	man-machine communication
Acoustic phonetics	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The 5 year, nine-contractor ARPA SUR project produced feasibility demonstrations of computer systems that understood spoken sentences related to restricted task domains. This report reviews and evaluates the contributions of that project, including the successful Harpy system developed at Carnegie-Mellon University (CMU) and other less successful but more ambitious systems developed at CMU, Bolt Beranek and Newman, and System Development Corporation. Harpy successfully understood 95% of the sentences spoken, for a task of document retrieval involving a 1000 word vocabulary. Performances of each system are discussed in light

over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

387 936

JCG

20.

of the complexities of the data retrieval or management tasks undertaken by the system. Major contributions to system design included Harpy's beam search technique and integrated network for representing phonetic, phonological, lexical, and syntactic informations about allowable sentences. Also, independent knowledge sources were organized in promising alternative forms in BBN's HWIM system and the CMU HEARSAY II system. Specific advances were made in procedures for vowel and consonant identification, phonological rules, lexical decoding, scoring procedures, applications of linguistic constraints, and in related research studies regarding phonetic and prosodic characteristics of spoken English sentences.

The project ~~is~~ was presented in a context of a history of 26 years of work on various forms of speech recognizers. Past and current technology is surveyed, including the currently available commercial products and several development projects. Gaps in current technology are defined and recommendations for filling those gaps are offered. Speech science centers and coordinated projects are suggested as means to advance the technology and relate the previous research to real applications.